

Fundamental Tradeoffs for Sparsity Pattern Recovery

Galen Reeves and Michael Gastpar

Department of Electrical Engineering and Computer Sciences

University of California, Berkeley

Email: {greeves, gastpar}@eecs.berkeley.edu

Abstract—Recovery of the sparsity pattern (or support) of a sparse vector from a small number of noisy linear samples is a common problem that arises in signal processing and statistics. In the high dimensional setting, it is known that recovery with a vanishing fraction of errors is impossible if the sampling rate and per-sample signal-to-noise ratio (SNR) are finite constants independent of the length of the vector. In this paper, it is shown that recovery with an arbitrarily small but constant fraction of errors is, however, possible, and that in some cases a computationally simple thresholding estimator is near-optimal. Upper bounds on the sampling rate needed to attain a desired fraction of errors are given in terms of the SNR and various key parameters of the unknown vector for two different estimators. The tightness of the bounds in a scaling sense, as a function of the SNR and the fraction of errors, is established by comparison with existing necessary bounds. Near optimality is shown for a wide variety of practically motivated signal models.

Index Terms—compressed sensing, information-theoretic bounds, random matrices, random projections, regression, sparse approximation, sparsity, subset selection.

I. INTRODUCTION

Recovery of sparse or compressible signals from a limited number of noisy linear projections is a problem that has received considerable attention in signal processing and statistics. Suppose, for instance, that a vector \mathbf{x} of length n is known to have exactly k nonzero elements, but the values and locations of these elements are unknown and must be estimated from a set of m noisy linear projections (or samples [1]) of the form

$$Y_i = \langle \phi_i, \mathbf{x} \rangle + W_i \quad \text{for } i = 1, \dots, m \quad (1)$$

where ϕ_i are known sampling vectors, $\langle \cdot, \cdot \rangle$ denotes the usual euclidean inner product, and W_i is additive white Gaussian noise. Then, a key insight from sparse signal recovery is that the number of samples required for reliable estimation depends primarily on the number of nonzero elements, and is potentially much less than the length of the vector \mathbf{x} .

One estimation problem of particular interest is to determine which elements of the vector \mathbf{x} are nonzero. This problem, which is referred to as *sparsity pattern* recovery in this paper, is known variously throughout the literature as support recovery or model selection and has applications in compressed sensing [2]–[4], sparse approximation [5], signal denoising [6], subset selection in regression [7], and structure estimation in graphical models [8].

A large body of recent work [8]–[20] has considered exact recovery of the sparsity pattern by deriving necessary and sufficient conditions on scalings of the tuple (n, k, m) to ensure that the probability of exact recovery tends to one as the vector length n becomes large. In particular, one line of work has considered the fundamental limitations of the recovery problem that apply to any possible estimator, regardless of computational complexity. In the noiseless setting, for instance, it has been shown that $m_n = k_n + 1$ samples are necessary and sufficient for an NP-hard combinatorial estimator [21], [22].

In presence of noise, however, support recovery depends critically on properties of the nonzero elements and cannot be characterized solely in terms of the dimensions n and k . In this setting, Wainwright [14] showed that $m = k + 1 + C \cdot k \log n$ samples are sufficient for an NP-hard combinatorial estimator where C is a finite constant that depends on the per-sample signal-to-noise ratio (SNR), the size of the smallest nonzero element, and various properties of the sampling vectors. Ensuing work by Fletcher, Rangan, and Goyal [18] and Wang, Wainwright, and Ramchandran [19], showed that, for a potentially different constant C , this scaling is also necessary for any algorithm.

In conjunction with the fundamental limitations outlined above, another line of work has studied scaling conditions for computationally tractable algorithms. In the noiseless setting, Donoho and Tanner [23] showed that $m = 2k \log(n/m)$ samples are sufficient for a polynomial-time linear program known as Basis Pursuit [6]. In the noisy setting, Wainwright [13] showed that $m = k + 1 + C_1 \cdot k \log n$ is sufficient for a polynomial-time algorithm known as the Lasso [24], and [18] showed that $m = k + 1 + C_2 \cdot k \log n$ is sufficient for a thresholding estimator where C_1 and C_2 are finite constants that depend on the SNR, the size of the smallest nonzero element, and various properties of the sampling vectors.

Although the scaling conditions outlined above show that exact recovery is possible with a relatively small number of samples, there exist two important limitations. First, if there is a non-vanishing fraction of nonzero elements, that is if $k/n \rightarrow \Omega$ for some *sparsity rate* $\Omega > 0$, then these results say that the ratio m/n must grow without bound with n in order to overcome the effect of noise. This behavior is in marked contrast to other recovery tasks, such as estimation of the vector \mathbf{x} with bounded mean squared error (MSE), which require only $m > \rho \cdot n$ samples for some finite *sampling rate* ρ .

(see [25], [26]). If noise is due to quantization, this means that accurate estimation with respect to MSE requires only fixed bit-rate whereas exact recovery of the sparsity pattern requires an unbounded bit-rate.

The second limitation is that scaling results in terms of the dimensions (n, k, m) do not tell the whole story. Often, one needs to know the exact constants involved in the bounds, or at least the dependence of these constants on parameters such as the SNR or various assumptions about the vector \mathbf{x} . For many of the estimation tasks considered throughout the compressed sensing literature, these properties are not well understood. As a result, the majority of sufficient conditions are far more conservative than those suggested by empirical evidence, and the optimality (or gap from optimality) of existing algorithms is difficult to determine due to the potential looseness of the necessary conditions.

A. Outline of Main results

In the present work, we derive upper bounds on the number of samples needed for *approximate* recovery of the sparsity pattern in the high dimensional setting when there exists a non-vanishing fraction of nonzero elements. In particular, we consider two different estimation algorithms—the NP-hard combinatorial optimization algorithm studied by Wainwright and the computationally efficient thresholding algorithm studied by Fletcher et al.—and characterize the sampling rate $\rho = m/n$ needed to ensure the number of errors in the estimated sparsity pattern does not exceed $\alpha \cdot k$ for some error rate α . Corresponding lower bounds, which apply to any estimator, are derived in the companion paper [20]. An example of these bounds is shown in Figure 1.

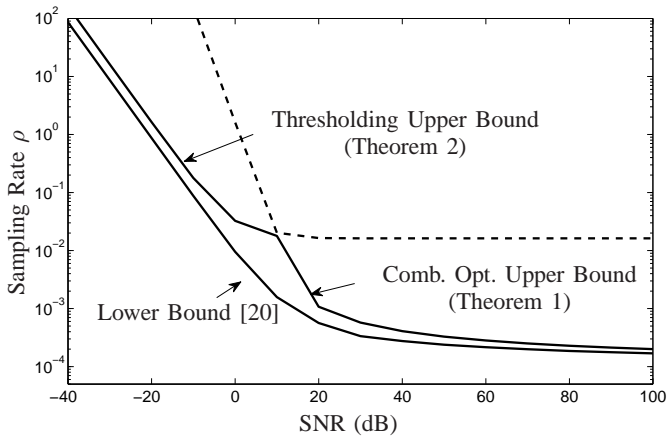


Fig. 1. Bounds on the asymptotic sampling rate $\rho = m/n$ and SNR required to identify the locations of at least 90% of the nonzero elements of a vector $\mathbf{x} \in \mathbb{R}^n$ with sparsity rate $\Omega = k/n = 10^{-4}$ when the power of the smallest nonzero element is at least 20% of the average power of the nonzero elements.

The contributions of this paper directly address the limitations of the scaling results for exact recovery outlined above. With respect to the first limitation, we show that any error fraction $\alpha > 0$ can be achieved using a finite sampling rate ρ . In other words, approximate sparsity pattern recovery has the same scaling behavior as estimation of \mathbf{x} with bounded mean squared error. If noise is due to quantization, this means that

a fixed bit-rate is sufficient for approximate sparsity pattern recovery.

With respect to the second limitation, our lower bounds are derived with an explicit dependence on various key problem parameters such as the SNR, the sparsity rate, and the relative size of the smallest nonzero elements. These bounds allow us to consider a wide variety of problem settings where the unknown vectors may be deterministic or stochastic and the magnitude of the smallest nonzero element may tend to zero as the vector length becomes large. Our framework allows us to address a number of important questions:

- What is the tradeoff between the sampling rate and the SNR? Our bounds show that there are two fundamentally different settings. If the SNR is small relative to the desired distortion then the number of samples needed is inversely proportional to the SNR. However, if the SNR is large relative to the desired distortion, then the number of samples needed is inversely proportional to the logarithm of the SNR.
- What is the tradeoff between optimality and computational complexity? As illustrated in Figure 1, our bounds show that a computationally simple thresholding estimator is near-optimal in the low-SNR setting. In the high SNR-setting, however, only the computationally hard estimator is shown to attain near-optimal performance. These results suggest that significant computational challenges arise in the high-SNR setting where the difficulty of estimation is due primarily to the uncertainty about the nonzero values.
- What happens as the desired error rate tends to zero? Our bounds show that the sampling rate depends on the inverse of the error rate $1/\alpha$. If the magnitudes of the nonzero elements have a fixed lower bound that is independent of n then this dependence is logarithmic. Otherwise, the dependence is polynomial.
- What is the effect of prior information? The upper bounds in this paper correspond to estimators that know the approximate number of nonzero elements, but have no prior information about their values. The lower bounds in the companion paper [20] apply to settings where the estimator may know statistical information such as the average power, range of values, or distribution. Interestingly, the resulting bounds show that in many cases this additional knowledge does not significantly improve the ability to estimate the sparsity pattern.

Beyond these results, our framework also permits us to prove some further insights. For instance, we show that the sampling rate distortion function is a convex function and that, in certain settings, i.i.d. sampling matrices are asymptotically strictly suboptimal.

This paper is organized as follows: Section II provides a precise problem formulation. Sections III and IV provide achievable bounds for two different estimators. Section V provides improved bounds using a particular set of sampling matrices that we will refer to as “rate-sharing” matrices. Section VI analyzes the scaling behavior of these bounds with respect to various key properties. Section VII presents

specific examples and illustrations, and proofs are given in the Appendices. The following section provides a brief, and necessarily incomplete, overview of work related to this paper.

B. Related Work

One line of related research [2]–[4], [6], [23]–[34] has focused on the approximation of sparse vectors with respect to the ℓ_2 norm. From a scaling perspective, one particularly important result from this literature [25], [26] is that any k -sparse vector \mathbf{x} of length n can be approximated with bounded mean squared error $\|\mathbf{x} - \hat{\mathbf{x}}\|^2/n < C_1/\text{SNR}$ using $m = \lceil C_2 \cdot k \log(n/k) \rceil$ samples and a quadratic program known as Basis Pursuit [6] where C_1, C_2 are finite constants. In the absence of any noise, this result provides sufficient conditions for exact recovery of \mathbf{x} , which in turn implies exact sparsity pattern recovery. In the presence of noise, however, it is important to observe that bounds on the mean squared error are insufficient to determine the accuracy of the estimated sparsity pattern.

Another line of related research [35]–[40] has focused on the rate-distortion behavior of sparse sources from an information-theoretic perspective. While these works provide valuable insights into the tradeoff between the SNR and the accuracy of approximation of \mathbf{x} in the ℓ_2 sense, they do not directly address the problem of sparsity pattern recovery.

Most closely related to the work in this paper is work which has focused directly on sparsity pattern recovery in the presence of noise [8]–[20]. As discussed in the introduction, it is now well understood that $m = k+1+C \cdot k \log n$ samples are both necessary and sufficient for exact recovery when the SNR is finite and there exists a fixed lower bound on the magnitude of the smallest nonzero elements.

The problem of approximate support recovery with a nonzero error rate α has also been considered in recent work. For the special case where the values of the nonzero elements are known, Aeron, Zhao, and Saligrama [11], [12] showed that $m = C \cdot k \log(n/k)$ samples are necessary and sufficient where the constant C is given explicitly in terms of the error rate α , the SNR, and nonzero values. In the more general setting where the nonzero values are unknown the necessary and sufficient condition $m = C \cdot k \log(n/k)$ was derived independently by Akcakaya and Tarokh [15] and the authors of this paper [16], [17], [20] for the special case where $k/n \rightarrow \Omega \in (0, 1)$. While the work of [15] provides bounds for a variety of scalings, actual upper and lower bounds on the constant C are not explicitly stated. By contrast, the upper bounds in this paper, in conjunction with the lower bounds in the companion paper [20], provide a tight characterization of the constant C for the setting of linear sparsity and show how this constant depends on various key properties such as the SNR, the size of the αk 'th smallest nonzero element, and the distortion α . Furthermore, while the upper bounds of [15] correspond to a computationally hard joint typicality decoding strategy, which requires knowledge of both the sparsity k as well as the SNR, we show the same scaling can be achieved using a computationally tractable thresholding estimator which depends only on the sparsity k .

II. PROBLEM FORMULATION

In this paper, we assume that \mathbf{x} is an arbitrary (non-random) element from some subset $\mathcal{X}^n \subset \mathbb{R}^n$. The *sparsity pattern* $\mathbf{s} \subseteq \mathcal{S}^n = \{1, 2, \dots, n\}$ is the set of integers indexing the nonzero elements of \mathbf{x} ,

$$\mathbf{s} := \{i : x_i \neq 0\},$$

and the sparsity $k = |\mathbf{s}|$ is the number of nonzero elements. We denote by \mathcal{S}_k^n the set of all subsets of \mathcal{S}^n of cardinality k . For simplicity, we assume that the sparsity k is known; in Section V we show that results obtained using this assumption can be extended to settings with only approximate knowledge about k .

We assume that \mathbf{x} is sampled using the noisy linear observation model given in (1). In matrix form, the vector of samples $\mathbf{Y} \in \mathbb{R}^m$ can be expressed as

$$\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known *sampling matrix* with rows equal to ϕ_i^T and $\mathbf{W} \in \mathbb{R}^m$ is a standard Gaussian vector. We further assume that an estimator (or recovery algorithm) is given the set $(\mathbf{Y}, \mathbf{A}, k)$, and the goal is to recover the sparsity pattern \mathbf{s} of \mathbf{x} .

To quantify the distortion between a sparsity pattern \mathbf{s} and its estimate $\hat{\mathbf{s}}$, it is important to observe that there are two different error events: one type of error occurs when an element in \mathbf{s} is omitted from the estimate $\hat{\mathbf{s}}$ and the other occurs when an element not present in \mathbf{s} is included in $\hat{\mathbf{s}}$. In this paper, we use the distortion function $d : \mathcal{S}^n \times \mathcal{S}^n \mapsto [0, 1]$ defined by

$$d(\mathbf{s}, \hat{\mathbf{s}}) := 1 - \frac{|\mathbf{s} \cap \hat{\mathbf{s}}|}{\max(|\mathbf{s}|, |\hat{\mathbf{s}}|)} \quad (2)$$

which corresponds to the maximum of the two types of errors. It can be verified that this distortion function is a metric on \mathcal{S}^n . We say that recovery is successful with respect to distortion $\alpha \in [0, 1]$ if $d(\mathbf{s}, \hat{\mathbf{s}}) \leq \alpha$. Exact recovery corresponds to the case $\alpha = 0$.

We are interested in performance guarantees that hold uniformly for any $\mathbf{x} \in \mathcal{X}^n$. It is important to note, however, that for any particular sampling matrix \mathbf{A} , there may exist a degenerate subset of \mathcal{X}^n for which recovery is particularly difficult. To overcome the effects of these sets, we allow \mathbf{A} to be a random matrix (denoted using boldface) distributed independently of \mathbf{x} . Given any sparsity pattern estimator $\hat{\mathbf{s}} : \mathbb{R}^m \times \mathbb{R}^{m \times n} \times \mathbb{N} \mapsto \mathcal{S}^n$, the probability of error corresponds to the worst case $\mathbf{x} \in \mathcal{X}^n$ with respect to the distribution on \mathbf{A} ,

$$P_e^{(n)} = \inf_{\mathbf{x} \in \mathcal{X}^n} \Pr \left\{ d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, k)) > \alpha \right\}.$$

Estimation in the presence of noise depends critically on the size of the entries in the sampling matrix. In this paper, we assume that

$$\mathbb{E}[\text{tr}(\mathbf{A}\mathbf{A}^T)] = m. \quad (3)$$

This scaling is consistent with the related work [3], [4], [12], [16], [17] and corresponds to the setting where each sampling

vector (i.e. row of \mathbf{A}) has unit magnitude. Thus, one useful property of this scaling is that the SNR of the linear samples given in (1) can be compared directly with that of classical samples of the form $Y_i = x_i + W_i$. Another useful property is that the SNR does not depend on the number of samples m .

We caution the reader that various other scalings of the sampling matrix are also used in the literature, and thus extra care is needed when comparing results. For instance, in [13], [14], [19] each element of \mathbf{A} has unit power, and the squared magnitude of each sampling vector is thus proportional to the vector length n .

To characterize the number of samples that are needed, we consider the high dimensional setting where the vector length n becomes large. We use \mathcal{X} to denote a sequence of subsets $\{\mathcal{X}^n : n \in \mathbb{N}\}$ and refer to \mathcal{X} as a *vector source*. The main question we address is whether or not recovery is possible when the number of samples is given by $m_n = \lceil \rho \cdot n \rceil$ for some finite *sampling rate* ρ that is a fixed constant independent of n . We use the notation $\hat{\mathbf{s}}$ interchangeably to denote an estimate of the sparsity pattern \mathbf{s} , or a family of estimators $\{\hat{\mathbf{s}}_{n,m} : n, m \in \mathbb{N}\}$.

Definition 1. A sampling rate distortion pair (ρ, α) is said to be *achievable* for a vector source \mathcal{X} if for each integer n there exists an estimator $\hat{\mathbf{s}}$ and $\lceil \rho \cdot n \rceil \times n$ sampling matrix \mathbf{A} such that

$$P_e^{(n)} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

The *sampling rate distortion function* $\rho(\alpha)$ of a vector source \mathcal{X} is the infimum of rates $\rho \geq 0$ such that the pair (ρ, α) is achievable. The *operational sampling rate distortion function* $\rho^{\text{opr}}(\alpha)$ of an estimator $\hat{\mathbf{s}}$ is the infimum of rates that are achievable using $\hat{\mathbf{s}}$.

We focus exclusively on the scaling regime where the sparsity k scales linearly with the vector length n .

Definition 2. Given any *sparsity rate* $\Omega \in (0, 1/2)$, the vector source $\mathcal{X}(\Omega)$ is the set of all sequences $\{\mathbf{x}^{(n)} \in \mathbb{R}^n : n \in \mathbb{N}\}$ for which $k_n/n \rightarrow \Omega$ as $n \rightarrow \infty$.

From a sampling perspective, the sparsity rate Ω measures the degrees of freedom per dimension of \mathbf{x} and is analogous to the rate of innovation [41] or “bandwidth” of an infinite length discrete time sequence.

One limitation of the general source $\mathcal{X}(\Omega)$ is that the nonzero values may be arbitrarily small, thus making recovery in the presence of noise impossible. In previous work [13], [14], this issue is addressed by placing a lower bound on the magnitudes of the nonzero elements of \mathbf{x} . This paper uses the more general approach where the nonzero elements are characterized by a distribution F or set of distribution functions \mathcal{F} . For any $\mathbf{x} \in \mathbb{R}^n$ and sparsity pattern \mathbf{s} we define

$$F_n(x) := \frac{1}{|\mathbf{s}|} \sum_{i \in \mathbf{s}} \mathbf{1}(x_i \leq x).$$

to be the empirical distribution of $\{x_i : i \in \mathbf{s}\}$.

Definition 3. Given any sparsity rate $\Omega \in (0, 1/2)$ and distribution F with finite second moment and zero probability

mass at zero, the vector source $\mathcal{X}(\Omega, F)$ is the set of all sequences $\{\mathbf{x}^n \in \mathbb{R}^n : n \in \mathbb{N}\}$ for which $k_n/n \rightarrow \Omega$ and $\|F_n - F\|_\infty \rightarrow 0$ as $n \rightarrow \infty$. Given any set of distributions \mathcal{F} , the vector source $\mathcal{X}(\Omega, \mathcal{F})$ consists of the union $\cup_{F \in \mathcal{F}} \mathcal{X}(\Omega, F)$.

To be consistent with previous work, we may for example consider the source $\mathcal{X}(\Omega, \mathcal{F})$ where \mathcal{F} denotes the set of all distributions whose support is bounded away from zero. However, one advantage of our approach is that we may also consider a source $\mathcal{X}(\Omega, F)$ where F has a density around zero, and thus a small number of nonzero elements may be arbitrarily small.

III. BOUNDS FOR COMBINATORIAL OPTIMIZATION

To understand the fundamental tradeoffs involved in sparsity pattern recovery, it is useful to consider the performance that can be achieved without any constraints on computational complexity. In this section, we study the behavior of a particular estimator which uses no prior information about the vector \mathbf{x} other than the number k of nonzero elements, but requires solving an NP-hard combinatorial optimization problem. This estimator, which was used by Wainwright [14] to give fundamental scaling results for exact recovery, corresponds to the maximum likelihood estimate of the vector \mathbf{x} or equivalently the least squares estimate over the ℓ_0 ball of size k . In this paper, we refer to this estimator as the *nearest subspace* estimator.

Definition 4 (Nearest Subspace Estimator). For a given set (\mathbf{y}, A, k) , the *nearest subspace* (NS) sparsity pattern estimate $\hat{\mathbf{S}}^{\text{NS}}$ is selected uniformly at random from the set

$$\arg \min_{\mathbf{s} \in \mathcal{S}_k^n} \|\Pi(A_{\mathbf{s}})\mathbf{y}\| \quad (4)$$

where $\Pi(A_{\mathbf{s}})$ denotes the $m \times m$ orthogonal projection matrix onto the null space of the $m \times k$ submatrix $A_{\mathbf{s}}$.

We remark that in many cases, the minimizer of (4) is unique and the nearest subspace estimate is a deterministic function of its inputs. In this section, for example, we derive bounds by considering distributions on the sampling matrix \mathbf{A} that guarantee uniqueness almost surely. In other important cases however, such as those considered in Section V, the minimizing set contains multiple sparsity patterns and performance guarantees require randomness.

Also, we remark that the nearest subspace estimator does not use any information about the nonzero values of \mathbf{x} . On one hand, this means that the performance of nearest subspace estimator for a source $\mathcal{X}(\Omega, F)$ may be suboptimal in general. (For the special case of Gaussian sources, a connection to optimal estimation in the high SNR setting is discussed in Section VIII.) On the other hand, this means that the operational sampling rate distortion function of a general source $\mathcal{X}(\Omega, \mathcal{F})$ corresponds directly to the worst case $F \in \mathcal{F}$, a useful property that is not necessarily true for estimators that depends on F .

Before we state our main results, we define the following the key properties of the vector source $\mathcal{X}(\Omega, F)$.

Definition 5. The power of a vector source $\mathcal{X}(\Omega, F)$ is given by

$$P(\Omega, F) = \Omega(\mu_F^2 + \sigma_F^2) \quad (5)$$

where μ_F and σ_F^2 are the mean and variance of F .

Due to the scaling of the sampling matrix given by (3), the power $P(\Omega, F)$ represents the SNR of the samples, that is

$$P(\Omega, F) = \lim_{n \rightarrow \infty} \frac{\mathbb{E}\|\mathbf{A}\mathbf{x}\|^2}{\mathbb{E}\|\mathbf{W}\|^2}.$$

Definition 6. For any $0 \leq \beta \leq 1$, the β -truncated distribution F_β of a distribution F is defined as

$$F_\beta(x) := \Pr\{X \leq x | Z = 0\} \quad (6)$$

where X has distribution F and $Z \in \{0, 1\}$ obeys

$$\Pr\{Z = 1\} = \begin{cases} 1, & \text{if } |X| > t_\beta \\ p_\beta, & \text{if } |X| = t_\beta \\ 0, & \text{if } |X| < t_\beta \end{cases}$$

with t_β and p_β chosen such that $\Pr\{Z = 0\} = \beta$.

The β -truncated distribution F_β characterizes the distribution of the smallest (in magnitude) βk nonzero elements of \mathbf{x} . For instance, if $F(x)$ has a nonzero density that is flat in a neighborhood around $x = 0$ then F_β converges to a uniform distribution as $\beta \rightarrow 0$.

Using the above definitions, we are able to state our main result which is an upper bound on the sampling rate distortion function $\rho(\alpha)$. The proof is given in Appendix A.

Theorem 1. The operational sampling rate distortion function $\rho^{\text{NS}}(\alpha)$ of the vector source $\mathcal{X}(\Omega, F)$ corresponding to the nearest subspace estimator is upper bounded by

$$\rho^{\text{NS-UB}}(\alpha) = \Omega + \max_{\alpha \leq \beta \leq 1} \frac{\Omega H(\beta) + (1 - \Omega)H(\frac{\Omega\beta}{1-\Omega})}{\mathcal{L}(1 + P(\beta\Omega, F_\beta))} \quad (7)$$

where $H(x) = -x \log(x) - (1 - x) \log(1 - x)$ denotes binary entropy,

$$\mathcal{L}(x) = \frac{1}{2} \left[\log(x) - \frac{x-1}{x} \right], \quad (8)$$

and $P(\cdot, \cdot)$ and F_β are defined by (5) and (6) respectively. Moreover, if the sampling matrix is i.i.d. Gaussian, then for any sampling rate $\rho > \rho^{\text{NS-UB}}(\alpha)$, there exists a constant $C > 0$ such that the probability of error obeys

$$P_e^{(n)} \leq \exp(-C \cdot n).$$

One immediate consequence of Theorem 1 is that any distortion $\alpha > 0$ can be achieved using a finite sampling rate ρ . With respect to scalings of (n, k, m) , this particular result has been shown in earlier version of this work [16], by Akcaya and Tarokh [15] in terms of unspecified constants, and by Aeron et al. [11], [12] for the discrete valued vectors.

A key difference with respect to previous work, however, is that Theorem 1 provides an explicit upper bound on the value of the sampling rate generally for any distribution F . This characterization makes it possible to understand how the fundamental sampling rate distortion function $\rho(\alpha)$ depends

on the distortion α , the SNR, or other key properties of the source.

For instance, the first term on the right hand side of (7) corresponds to the noiseless sampling rate distortion function of the general source $\mathcal{X}(\Omega)$ when the sampling matrix is constrained to have i.i.d. elements (see [20, Proposition 1]). The second term corresponds to the additional sampling rate needed to overcome the noise. At high SNR, this term is inversely proportional to the logarithm of the SNR. Further analysis of the bound in Theorem 1 with respect to scalings of α and the SNR is provided in Section VI.

IV. BOUNDS FOR THRESHOLDING

One question of practical importance is whether there exist computationally efficient estimators whose recovery performance is comparable to that of computationally unconstrained estimators such as the nearest subspace estimator studied in the previous section. In this section, we bound the sampling rate distortion function of a particular thresholding-style estimator and show that in some cases, near-optimal performance can be attained. The estimator we study was first introduced by Fletcher et al [18], under the name *maximum correlation*, for the study of exact sparsity pattern recovery. In this paper, we refer to it as the *thresholding* estimator for reasons that will become clear shortly.

Definition 7 (Thresholding Estimator). For a given set (\mathbf{y}, A, k) , the *thresholding* (TH) sparsity pattern estimate $\hat{\mathbf{S}}^{\text{TH}}$ is selected uniformly at random from the set

$$\arg \max_{\mathbf{s} \in S_k^n} \|A_{\mathbf{s}}^T \mathbf{y}\|. \quad (9)$$

Although the optimization in (9) appears to be similar to the optimization problem in the nearest subspace estimate (4), the key difference is that the above problem can be solved efficiently by identifying the k largest elements of the n -dimensional $\mathbf{z} = A^T \mathbf{y}$. If the minimizer of (9) is unique, then the thresholding estimate can be expressed equivalently as

$$\hat{\mathbf{S}}^{\text{TH}} = \{i \in \{1, 2, \dots, n\} : |z_i| \geq t_k(\mathbf{z})\}$$

where $t_k(\mathbf{z})$ is the magnitude of the k 'th largest nonzero element of \mathbf{z} . A connection between the thresholding estimator and optimal estimation for Gaussian sources at low SNR is discussed further in Section VIII.

We now give our main result which is an exact characterization of the sampling rate distortion function $\rho(\alpha)$. The proof is given in Appendix B.

Theorem 2. The operational sampling rate distortion function $\rho^{\text{TH}}(\alpha)$ of the vector source $\mathcal{X}(\Omega, F)$ corresponding to the thresholding estimator is upper bounded by the solution $\rho^{\text{TH-UB}}(\alpha)$ to

$$\int_{\mathbb{R}} G\left(\frac{\rho^{\text{TH-UB}}(\alpha) x^2}{1 + P(\Omega, F)}, Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right)\right) dF(x) = \alpha \quad (10)$$

if $\alpha < 1 - \Omega$ and is zero otherwise where $G(\mu^2, t) = 1 - Q(t + \mu) - Q(t - \mu)$, $Q(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$ and $P(\cdot, \cdot)$ is defined by (5). Moreover, if the sampling matrix is i.i.d. Gaussian, then $\rho^{\text{TH}}(\alpha) = \rho^{\text{TH-UB}}(\alpha)$.

In the special case where the distribution F is a zero-mean Gaussian, the solution to (10) can be expressed explicitly.

Corollary 1. *If F is a zero-mean Gaussian distribution, then the upper bound on the operational sampling rate distortion function of the thresholding estimator given in Theorem 2 is given by*

$$\rho^{\text{TH-UB}}(\alpha) = \Omega \frac{1 + P(\Omega, F)}{P(\Omega, F)} \left[\left(\frac{Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right)}{Q^{-1}\left(\frac{\alpha}{2}\right)} \right)^2 - 1 \right]. \quad (11)$$

More generally, the solution to (10) can be upper bounded for any distribution F using the following result which depends only on the average power and the β -truncated distribution of F . The proof is given in Appendix B.

Corollary 2. *The upper bound on the operational sampling rate distortion function of the thresholding estimator given in Theorem 2 obeys*

$$\rho^{\text{TH-UB}}(\alpha) \leq \Omega \frac{1 + P(\Omega, F)}{P(\Omega, F_{\alpha/2})} \left[Q^{-1}\left(\frac{\alpha}{2}\right) + Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right) \right]^2 \quad (12)$$

where F_β is defined by (6).

A key contribution of Theorem 2 is that any distortion $\alpha > 0$ can be achieved using a finite sampling rate and a computationally efficient estimator. This means that, with respect to scalings of the dimensions (n, k, m) , the performance of the thresholding estimator is equivalent to that of the nearest subspace estimator.

A further contribution of Theorem 2 is that in many cases, the sampling rate distortion function $\rho^{\text{TH}}(\alpha)$ is significantly less than the upper bound for the nearest subspace estimator $\rho^{\text{NS-UB}}(\alpha)$ given in Theorem 1 and thus provides an improved upper bound on the fundamental sampling rate distortion function $\rho(\alpha)$. In Section VI, we use this improved upper bound to characterize the rate at which $\rho(\alpha)$ tends to infinity as α tends to zero and the rate at which $\rho(\alpha)$ tends to infinity as the SNR tends to zero.

At a high level, our proof of Theorem 2 is similar to the proof used by Fletcher et al. [18] for exact recovery in the sense that both proofs depend on the asymptotic behavior of the vector $\mathbf{A}^T \mathbf{x}$. Technically, however, there is a key difference: whereas exact recovery depends on the extreme order statistics, which can be controlled using a union bound, approximate recovery depends on the limiting empirical distribution. Since the elements of $\mathbf{A}^T \mathbf{x}$ are not independent, the main challenge in our proof is showing convergence of the empirical distribution. We use standard truncation arguments as well as Pinsker's inequality and manipulation of various mutual informations.

V. RATE-SHARING SAMPLING MATRICES

The upper bounds on the sampling rate distortion function $\rho(\alpha)$ in Theorems 1 and 2 were derived by analyzing the special case where the elements of \mathbf{A} are i.i.d. zero-mean Gaussian. In this section, we show that any upper bound on the

operational sampling rate distortion function $\rho^{\text{opr}}(\alpha)$ of a vector source $\mathcal{X}(\Omega, F)$ and estimator $\hat{\mathbf{s}}$ can be convexified using “rate-sharing” sampling matrices. This result both strengthens our previous bounds and proves that $\rho(\alpha)$ is a convex function.

Before we discuss this strategy, we need the following useful result which shows that the operational sampling rate distortion function $\rho^{\text{opr}}(\alpha)$ of an estimator $\hat{\mathbf{s}}$ does not change if the estimator only has approximate information about the sparsity k_n . The proof is given in Appendix C.

Theorem 3. *Let $\rho^{\text{opr}}(\alpha)$ be the operational sampling rate distortion function of a vector source $\mathcal{X}(\Omega, F)$ and estimator $\hat{\mathbf{s}}$. Let $\tilde{\rho}^{\text{opr}}(\alpha)$ be the corresponding operational sampling rate distortion function when the estimator $\hat{\mathbf{s}}$ uses some sparsity sequence \tilde{k}_n instead of the true sparsity sequence k_n . If $\lim_{n \rightarrow \infty} |\tilde{k}_n - k_n|/n = 0$ and $\limsup_{n \rightarrow \infty} k_n - \tilde{k}_n \leq 0$, then $\tilde{\rho}^{\text{opr}}(\alpha) = \rho^{\text{opr}}(\alpha)$.*

One consequence of Theorem 3 is that our bounds on $\rho(\alpha)$ apply to the setting where the number of elements is a random variable that concentrates around the expected value Ωn . This occurs, for example, if the elements of \mathbf{x} are i.i.d. random variables with distribution function F_X given by

$$F_X(x) = (1 - \Omega)\mathbf{1}(x = 0) + \Omega F(x).$$

The fact that the sparsity k_n does not need to be known exactly is also an important part of our convexification strategy which is described in the following result. The proof is given in Appendix D.

Theorem 4. *Let (ρ_1, α_1) and (ρ_2, α_2) be two achievable sampling rate distortion pairs for a vector source $\mathcal{X}(\Omega, F)$ and let $\{\mathbf{A}_1^{(n)}\} \in \mathbb{R}^{\lceil \rho_1 \cdot n \rceil \times n}$ and $\{\mathbf{A}_2^{(n)}\} \in \mathbb{R}^{\lceil \rho_2 \cdot n \rceil \times n}$ be the sampling matrix sequences that achieve these rates. For any $\lambda \in (0, 1)$, let $\{\mathbf{A}^{(n)}\}$ be a rate-sharing sampling matrix sequence defined by*

$$\mathbf{A}^{(n)} = \begin{bmatrix} \mathbf{A}_1^{(\lceil \lambda \cdot n \rceil)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2^{(n - \lceil \lambda \cdot n \rceil)} \end{bmatrix} \mathbf{P}^{(n)} \quad (13)$$

where $\mathbf{P}^{(n)}$ is distributed uniformly over the set of $n \times n$ permutation matrices. Then, the sampling rate distortion pair $(\lambda\rho_1 + (1 - \lambda)\rho_2, \lambda\alpha_1 + (1 - \lambda)\alpha_2)$ is achievable using the sequence $\{\mathbf{A}^{(n)}\}$.

Two consequences of Theorem 4 are the following.

Corollary 3. *For any estimator $\hat{\mathbf{s}}$, the operational sampling rate distortion function $\rho^{\text{opr}}(\alpha)$ of the vector source $\mathcal{X}(\Omega, F)$ is convex.*

Corollary 4. *The sampling rate distortion function $\rho(\alpha)$ of the vector source $\mathcal{X}(\Omega, F)$ is convex.*

Additionally, since $\rho(\alpha) = 0$ for any distortion $\alpha \geq 1 - \Omega$, we may conclude that

$$\rho(\lambda\alpha + (1 - \lambda)(1 - \Omega)) \leq \lambda\rho(\alpha) \quad (14)$$

for all $0 \leq \lambda \leq 1$.

From a practical standpoint, the rate-sharing strategy used in Theorem 4 shows that, in the high-dimensional setting, our

sampling problem can be split into two separate subproblems. Since the splitting is done randomly, the properties of each subproblem, such as the empirical distribution of \mathbf{x} , can be accurately characterized in terms of the original problem.

Lastly, we remark upon a key difference between the i.i.d. Gaussian matrices used in Theorems 1 and 2 and the rate-sharing matrices described in Theorem 4 with respect to recovering vectors that are sparse in a basis other than the standard basis. In particular, suppose that \mathbf{x} is not actually sparse, but instead has some sparse representation $\tilde{\mathbf{x}}$ with respect to an orthonormal matrix $B \in \mathbb{R}^{n \times n}$, that is $\mathbf{x} = B\tilde{\mathbf{x}}$. If both the designer of the sampling matrix \mathbf{A} and the estimator $\hat{\mathbf{s}}$ know B , then any matrix \mathbf{A} designed for the original setting (where $B = I_n$) can be realized in the generalized setting by using the modified sampling matrix $\mathbf{A}B^{-1}$. What happens, however, if only the estimator knows B ? Since any i.i.d. Gaussian matrix \mathbf{A} is equal in distribution to $\mathbf{A}B$, the bounds given in Theorems 1 and 2 still apply. By contrast, the rate-sharing matrix in Theorem 4 depends critically on the knowledge of B and cannot be used if \mathbf{A} must be designed independently of B . For a further discussion of this issue, see [42].

VI. SCALING BEHAVIOR

In this section, we show how the bounds on the sampling rate distortion function $\rho(\alpha)$ given in Theorems 1 and 2 depend on the distortion α and various properties of the source $\mathcal{X}(\Omega, F)$ such as the sparsity rate Ω or the power $P(\Omega, F)$. By comparing simplified versions of the upper bounds in this paper with the lower bounds from the companion paper [20], we are able address questions such as how $\rho(\alpha)$ increases as α becomes small and how $\rho(\alpha)$ converges to the noiseless rate $\rho_0(\alpha)$ as the SNR becomes large.

One key property of the source is the power $P(\Omega, F)$. To describe scalings of the power we use $\mathcal{X}(\Omega, F; P)$ to denote a source characterized by a distribution F that is scaled to have power P . Another key property of the source is the following.

Definition 8. The decay rate $L \in [0, \infty]$ of a distribution function F is defined as

$$L := \lim_{\epsilon \rightarrow 0} \frac{\log \epsilon}{\log (F(\epsilon) - F(-\epsilon))} \quad (15)$$

if the limit exists.

The decay rate L characterizes the relative size the smallest nonzero elements drawn from a source $\mathcal{X}(\Omega, F)$. For instance, if X is a random variable with decay rate $L < \infty$, and we define

$$x_\epsilon = \inf \{x \geq 0 : \Pr\{|X| \leq x\} \geq \epsilon\},$$

then $\epsilon^{-L} \cdot x_\epsilon \rightarrow c$ as $\epsilon \rightarrow 0$ for some $c \in (0, \infty)$. The decay rate is $L = 0$ if X is bounded away from zero and $L = \infty$ if and only if $\Pr\{X = 0\} > 0$. We denote by \mathcal{F}_0 the set of all distributions F with finite second moment and zero probability mass at zero, and note that L is finite for any $F \in \mathcal{F}_0$.

One useful property of the decay rate is that it can be used to bound the relative power of the β -truncated distribution given in Definition 6.

Lemma 1 ([20]). *Given any distribution function $F \in \mathcal{F}_0$ with decay rate L , there exist constants $0 < C_F^- \leq C_F^+ < \infty$ such that*

$$C_F^- \cdot \beta^{2L} \leq \frac{P(\Omega, F_\beta)}{P(\Omega, F)} \leq C_F^+ \cdot \beta^{2L} \quad (16)$$

for any $0 \leq \beta \leq 1$.

Using the above properties, we are able to provide the following simplified versions of Theorem 1 and 2.

Proposition 1. *Given any distribution $F \in \mathcal{F}_0$, there exists a constant $C_F < \infty$ such that the operational sampling rate distortion function $\rho^{\text{NS}}(\alpha)$ of the vector source $\mathcal{X}(\Omega, F; P)$ corresponding to the nearest subspace estimator is upper bounded by*

$$\rho^{\text{NS}}(\alpha) \leq \Omega + C_F \cdot \max_{\beta \in \{\alpha, 1\}} \frac{\beta \Omega \log(\frac{1}{\beta \Omega})}{\log(1 + \beta^{4L+2} P^2)} \quad (17)$$

for all distortions $\alpha \in (0, 1/4)$ where F has decay rate L .

Proposition 2. *Given any distribution $F \in \mathcal{F}_0$, there exists a constant $C_F < \infty$ such that the operational sampling rate distortion function $\rho^{\text{TH}}(\alpha)$ of the vector source $\mathcal{X}(\Omega, F; P)$ corresponding to the thresholding estimator is upper bounded by*

$$\rho^{\text{TH}}(\alpha) \leq C_F \cdot \left(\frac{1+P}{P} \right) \frac{\Omega \log(\frac{1}{\alpha \Omega})}{\alpha^{2L}} \quad (18)$$

for all distortions $\alpha \in (0, 1/4)$ where F has decay rate L .

To understand the significance of Propositions 1 and 2, it is useful to consider the lower bounds from [20].

Proposition 3 ([20]). *Given any distribution $F \in \mathcal{F}_0$, there exists a constant $C_F > 0$ such that the sampling rate distortion function $\rho(\alpha)$ of the vector source $\mathcal{X}(\Omega, F; P)$ is lower bounded by*

$$\rho(\alpha) \geq C_F \cdot \frac{\alpha \Omega \log(\frac{1}{\alpha \Omega})}{\log(1 + \alpha^{2L+1} P)} \quad (19)$$

for all distortions $\alpha \in (0, 1/4)$ where F has decay rate L .

Combining Propositions 2 and 3 provides the following tight characterization of the scaling of $\rho(\alpha)$ with respect to α . (This result corresponds to Proposition 8 in [20].)

Proposition 4. *Given any distribution $F \in \mathcal{F}_0$ and sparsity rate Ω , there exist constants $0 < C_{F,\Omega}^- \leq C_{F,\Omega}^+ < \infty$ such that the sampling rate distortion function $\rho(\alpha)$ of the vector source $\mathcal{X}(\Omega, F)$ obeys*

$$C_{F,\Omega}^- \left(\frac{1}{\alpha} \right)^{-2L} \log \left(\frac{1}{\alpha} \right) \leq \rho(\alpha) \leq C_{F,\Omega}^+ \left(\frac{1}{\alpha} \right)^{-2L} \log \left(\frac{1}{\alpha} \right) \quad (20)$$

for all distortions $\alpha \in (0, 1/4)$ where F has decay rate L .

Furthermore, combining Propositions 1, 2, and 3 gives the following characterization of the scaling of $\rho(\alpha)$ with respect to the SNR.

Proposition 5. *Given any distribution $F \in \mathcal{F}_0$, sparsity rate Ω , and distortion $\alpha \in (0, 1/4)$, there exist constants $0 <$*

$C_{F,\Omega,\alpha}^- \leq C_{F,\Omega,\alpha}^+ < \infty$ such that the sampling rate distortion function $\rho(\alpha)$ of the vector source $\mathcal{X}(\Omega, F; P)$ obeys

$$\frac{C_{F,\Omega,\alpha}^-}{\log(1+P)} \leq \rho(\alpha) \leq \Omega + \frac{C_{F,\Omega,\alpha}^+}{\log(1+P)} \quad (21)$$

Proposition 5 shows that the tradeoff between the sampling rate ρ and the SNR exhibits two different behaviors: at low SNR, the sampling rate is inversely proportional to the SNR and at high SNR, the sampling rate is inversely proportional to the logarithm of the SNR.

The constant gap between the bounds in Proposition 5, corresponds, roughly speaking, to the use of trivial bounds $0 \leq \rho_0(\alpha) \leq \Omega$ on the noiseless sampling rate distortion function $\rho_0(\alpha)$. In some cases (see for example Proposition 2 in [20]) the lower bound is tight. However, in this special case where the sampling matrix is constrained to have i.i.d. elements, we may use stronger bounds from [20] to show that the upper bound is tight. To state this result, we need the following function which measures the relative entropy power of the distribution F . Given any sparsity rate Ω and distribution F with mean μ_F , variance σ_F^2 , and differential entropy $h(F)$, we define the function $\theta(\Omega, F) \in [0, 1]$ to be

$$\theta(\Omega, F) = \frac{(2\pi e)^{-1} \exp(2h(F))}{\sigma_F^2 + (1-\Omega)\mu_F^2}. \quad (22)$$

Proposition 6 ([20]). *If the sampling matrix is i.i.d., then there exists a constant $C > 0$ such that the sampling rate distortion function $\rho(\alpha)$ of the vector source $\mathcal{X}(\Omega, F; P)$ is lower bounded by*

$$\rho(\alpha) \geq \Omega + C \cdot \frac{\Omega \log(\frac{1}{\Omega})}{\log(1+P)} \quad (23)$$

for all distortions $\alpha \in (0, 1/4)$ that satisfy

$$\theta(\Omega, F) > \exp\left(1 - \frac{1}{\Omega} R(\Omega, \alpha)\right) \quad (24)$$

where $R(\Omega, \alpha) = H(\Omega) - \Omega H(\alpha) - (1-\Omega)H(\frac{\Omega\alpha}{1-\Omega})$.

VII. EXAMPLES AND ILLUSTRATIONS

This section provides specific examples and illustrations of the upper bounds on the sampling rate distortion function $\rho(\alpha)$ given in Theorems 1, 2, and 4.

A. Bounds for a Gaussian Source

To begin, we consider the setting where F is the distribution of a zero-mean Gaussian random variable with variance σ^2 . It is clear to see that the power of the source $\mathcal{X}(\Omega, F)$ is given by $P(\Omega, F) = \Omega\sigma^2$, and it can be shown (see Appendix D of [20]) that the power corresponding to the β -truncated distribution F_β is given by

$$P(\Omega, F_\beta) = \left[1 - (t_\beta/\beta)(2/\pi)^{1/2} \exp(-t_\beta^2/2)\right] \Omega \sigma_F^2 \quad (25)$$

where $t_\beta = Q^{-1}(\frac{1-\beta}{2})$ and $Q(\cdot)^{-1}$ denotes the functional inverse of $Q(x) = \int_x^\infty (2\pi)^{-1/2} \exp(-x^2/2) dx$.

Since the Gaussian distribution has decay rate $L = 1$, we know that the power $P(\Omega, F_\beta)$ scales like β^2 for small β .

Applying a Taylor expansion to the to (25) gives the more precise characterization

$$\lim_{\beta \rightarrow 0} \frac{1}{\beta^2} P(\Omega, F_\beta) = \frac{\pi}{6} \Omega.$$

To illustrate our bounds on the sampling rate distortion function of this source, we first consider the special case where the sampling matrix is required to have i.i.d. elements. In Figure 2(a)-(b), the upper bounds on $\rho(\alpha)$ given in Theorems 1 and 2 are shown as a function of α for two different SNRs. Since these bounds correspond to an i.i.d. Gaussian sampling matrix, they represent valid upper bounds for i.i.d. sampling matrices. Also shown is a lower bound (see [20, Theorem 4]) which applies universally to any estimator and i.i.d. sampling matrix.

In the low SNR setting (Figure 2(a)) the best bound corresponds to the thresholding estimator, and in the high SNR setting (Figure 2(b)) the best bound corresponds to the nearest subspace estimator. Since our analysis of the thresholding estimator is exact (at least for the case of a Gaussian sampling matrix), these results indicate that the nearest subspace estimator is significantly better than the thresholding estimator at high SNR. However, since the upper bound on the nearest subspace estimator is not necessarily tight, it is not possible to say which estimator is better at low SNR. It is interesting to observe that in both cases, the bounds show that $\rho(\alpha)$ is relatively flat for nonzero distortions α , but tend to infinity rapidly as $\alpha \rightarrow 0$. This behavior shows why it is important to consider approximate recovery of the sparsity pattern.

Next, we address what happens if there are no constraints on the sampling matrix (other than the normalization imposed by (3)). In Figure 3(a)-(b), convex versions of the upper bound on $\rho(\alpha)$ given in Theorems 1 and 2 corresponding to the rate-sharing strategy given in Theorem 4 are shown as a function of α for two different SNRs. Also shown is a lower bound (see [20, Theorem 3]) which applies universally to any estimator and sampling matrix.

The basic behavior of the bounds in Figure 3 is similar to that of the bounds for i.i.d. sampling matrices in Figure 2. For the upper bounds, the main difference between the two settings occurs at relatively large distortions. For the lower bounds, the difference is most prominent at high SNR. It is particularly interesting to note that for large distortions, the high SNR nearest subspace bound in Figure 3(b) is less than the corresponding lower bound in Figure 2(b). This behavior addresses the question as to whether or not i.i.d. sampling matrices are optimal in the asymptotic setting and shows that in certain cases such matrices are suboptimal.

B. Bounds for a General Source

We now show how our upper bounds, which are stated in terms of a single distribution function F , can be extended to more general sources characterized by a set of distributions \mathcal{F} . In particular, we consider the source $\mathcal{X}(\Omega, \mathcal{F}(\eta, \gamma))$ where, for any parameters $\eta \in [0, 1]$ and $\gamma \in (0, \infty)$, we define $\mathcal{F}(\eta, \gamma) \subset \mathcal{F}_0$ to be the set of all distributions with power γ and a lower bound $\sqrt{\eta\gamma}$ on the magnitude of any realization. This source, which we refer to as the *bounded source*, assumes

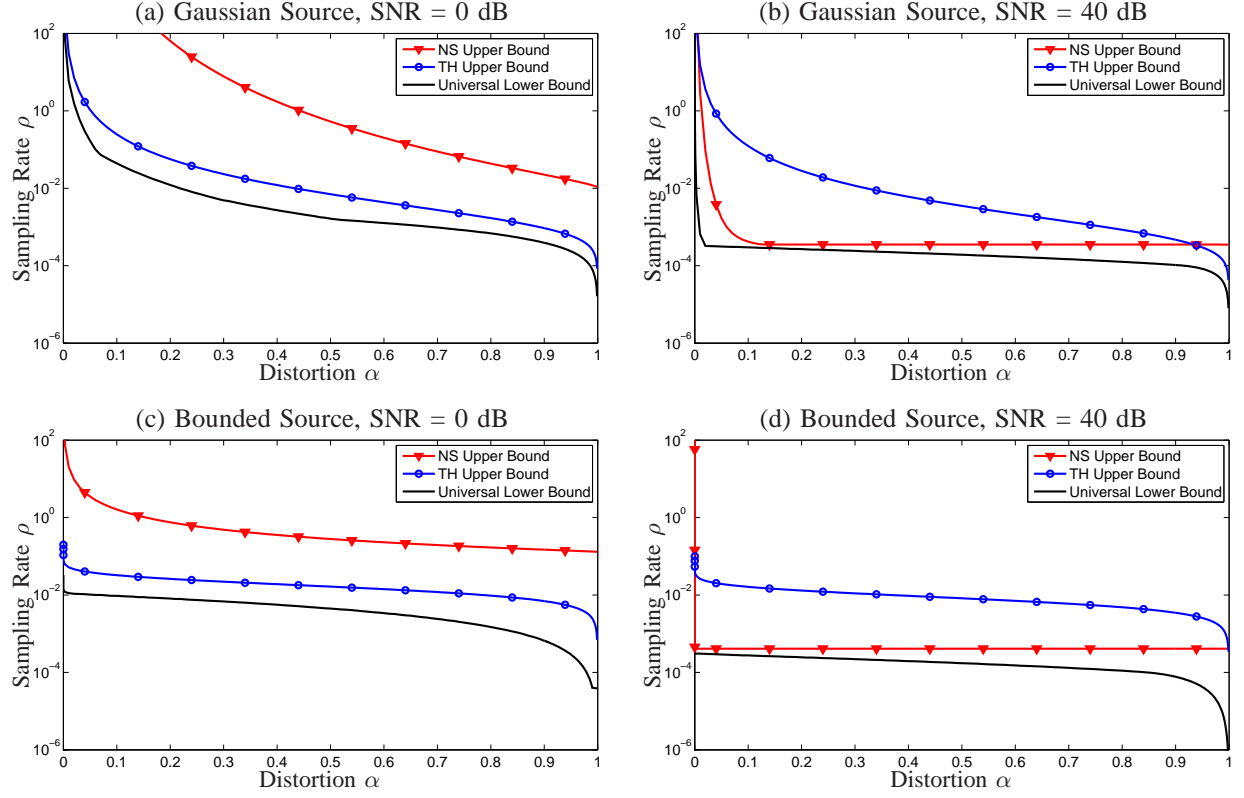


Fig. 2. Bounds of sampling rate distortion function $\rho(\alpha)$ for sampling matrices with i.i.d. elements. The upper bounds correspond to Theorems 1 and 2. The lower bound corresponds to Theorem 4 in [20]. Plots (a)-(b) correspond to a zero-mean Gaussian source. Plots (c)-(d) correspond to the general bounded source $\mathcal{X}(\Omega, \mathcal{F}(\eta, \gamma))$ with $\eta = 0.2$. In all cases $\Omega = 10^{-4}$.

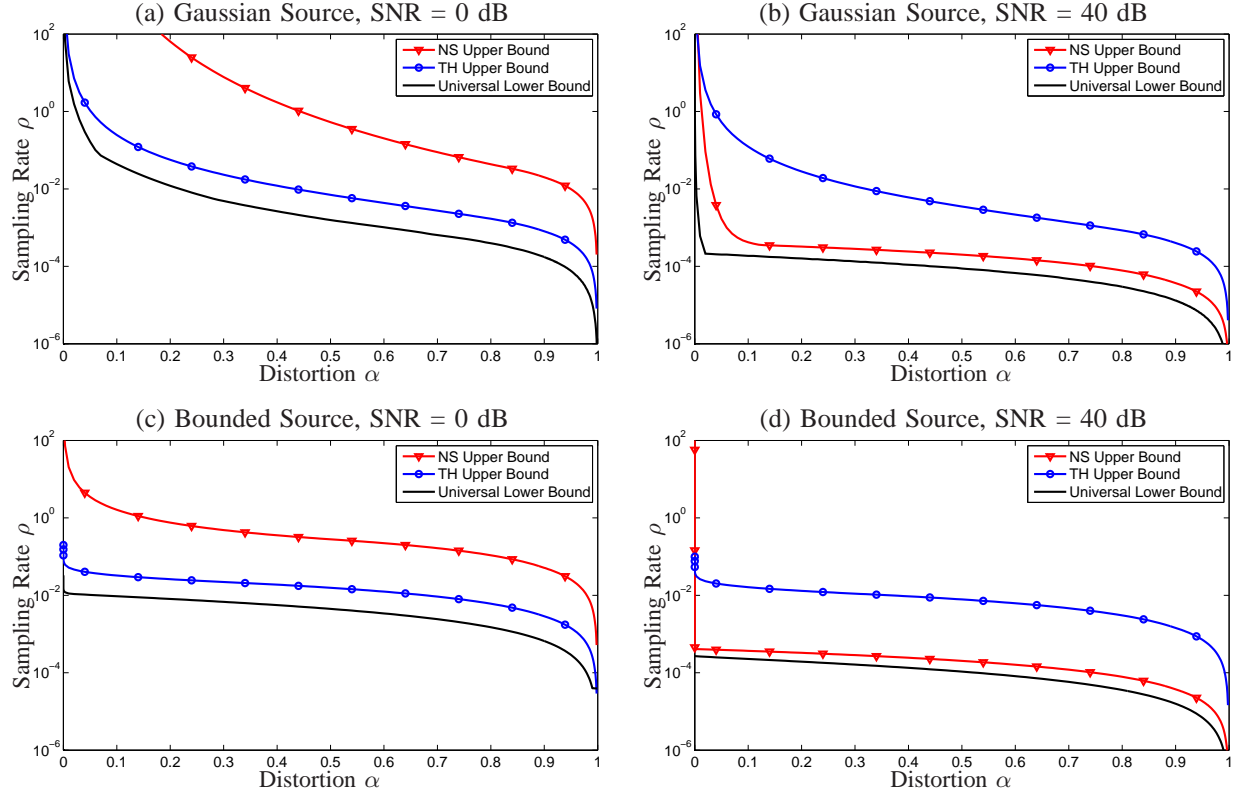


Fig. 3. Bounds of sampling rate distortion function $\rho(\alpha)$ for optimal sampling matrices. The upper bounds correspond to the convex versions of Theorems 1 and 2 that are attained using the rate-sharing sampling matrices described Theorem 4. The lower bound corresponds to Theorem 3 in [20]. Plots (a)-(b) correspond to a zero-mean Gaussian source. Plots (c)-(d) correspond to the general bounded source $\mathcal{X}(\Omega, \mathcal{F}(\eta, \gamma))$ with $\eta = 0.2$. In all cases $\Omega = 10^{-4}$.

very little about the actual nonzero values and corresponds to assumptions commonly used in literature for exact recovery [8], [9], [11]–[20].

To apply the bounds given in Theorems 1 and 2 we use the simple fact that, for any set of distributions \mathcal{F} and fixed estimator \hat{s} , the operational sampling rate distortion function of the source $\mathcal{X}(\Omega, \mathcal{F})$ is equal to the supremum over $F \in \mathcal{F}$ of the operational sampling rate distortion of the source $\mathcal{X}(\Omega, F)$. In other words, we may uniformly bound the source $\mathcal{X}(\Omega, \mathcal{F})$ by considering the worst case distribution $F \in \mathcal{F}$. We remark that this useful property, which is true for any fixed estimator \hat{s} , is not necessarily true for the fundamental sampling rate distortion function since the optimal estimator for a source $\mathcal{X}(\Omega, F)$ depends on the distribution F .

To bound the operational sampling rate distortion function corresponding to the nearest subspace estimator, we observe that for any distribution F , a lower bound on the power $P(\Omega, F_\beta)$ of the β -truncated distribution F_β provides an upper bound on the right hand side of (7). Using the fact that $\inf_{F \in \mathcal{F}(\eta, \gamma)} P(\Omega, F_\beta) = \beta\Omega\eta\gamma$, gives the following upper bound

$$\rho^{\text{NS-UB}}(\alpha) = \Omega + \max_{\alpha \leq \beta \leq 1} \frac{\Omega H(\beta) + (1-\Omega)H(\frac{\Omega\beta}{1-\Omega})}{\mathcal{L}(1 + \beta\Omega\eta\gamma)}. \quad (26)$$

To bound the operational sampling rate distortion function corresponding to the thresholding estimator, we use Corollary 2 and the fact that

$$\sup_{F \in \mathcal{F}(\eta, \gamma)} \frac{1 + P(\Omega, F)}{P(\Omega, F_{\alpha/2})} = \frac{1 + \Omega\gamma}{\eta\gamma\Omega}$$

to obtain the upper bound

$$\rho^{\text{TH-UB}}(\alpha) \leq \Omega \frac{1 + \Omega\gamma}{\eta\gamma\Omega} \left[Q^{-1}(\frac{\alpha}{2}) + Q^{-1}(\frac{\alpha\Omega}{2(1-\Omega)}) \right]^2. \quad (27)$$

In Figure 2(c)-(d), the upper bounds (26) and (27) are shown as a function of the the distortion α for two different SNRs. Also shown is a lower bound designed specifically for the bounded source $\mathcal{X}(\Omega, \mathcal{F}(\eta, \gamma))$ and i.i.d. sampling matrices (see [20, Section VI-C]). In Figure 1, the same bounds are shown as a function of the SNR for fixed α . In Figure 3 (c)-(d), convex versions of the upper bounds (26) and (27) corresponding to the rate-sharing strategy given in Theorem 4 are shown as a function of α for various SNRs. Also shown is a lower bound (see [20, Theorem 3]) which applies universally to any estimator and sampling matrix.

Our bounds for the bounded source have many similarities with the bounds for the Gaussian source. In both cases, the nearest subspace estimator bound is stronger at high SNR and the thresholding bounds is stronger at low SNR. Also, in both cases, the bounds are relatively flat for a range of α , but increase rapidly as α tends to zero. The main difference between bounded source and Gaussian source, however, is the point at which this change in behavior occurs. For example, in Figures 3(a)-(b) the bounds increase rapidly for all α less than 0.05. In the corresponding plots for the bounded source, Figure 3(c)-(d), this same behavior occurs, but only for α much less than 0.01.

The reason that small distortions are easier to obtain for the bounded source than for the Gaussian source is that the

difficulty of recovery for $\alpha \approx 0$ is dominated by the size of the smallest nonzero values. For the bounded source, these values are bounded away from zero whereas for the Gaussian source the values may be arbitrarily small. This property of a source is precisely what the decay rate L given Definition 8 is designed to capture. For example, using the fact that the Gaussian distribution has decay rate $L = 1$ and every distribution $F \in \mathcal{F}(\eta, \gamma)$ has decay rate $L = 0$, the precise rate at which $\rho(\alpha)$ increases as $\alpha \rightarrow 0$ can be determined using Proposition 4.

VIII. CONNECTIONS WITH OPTIMAL ESTIMATION

In this last section, we study the optimal estimator for the vector source $\mathcal{X}(\Omega, F)$ characterized by a zero-mean Gaussian distribution F , and show that certain limiting versions of this estimator correspond to the nearest subspace and thresholding estimators studied in Sections III and IV.

Our first result describes the optimal estimator associated with our distortion metric (2). The proof, which follows directly from the Bayesian formulation of the problem, is left as an exercise.

Proposition 7. *Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector whose sparsity pattern \mathbf{S} is distributed uniformly over \mathcal{S}_k^n and whose nonzero elements $\{X_i : i \in \mathbf{s}\}$ are standard Gaussian random variables, and let*

$$\mathbf{Y} = \sqrt{\gamma} \cdot A\mathbf{X} + \sqrt{1-\gamma} \cdot \mathbf{W}$$

for some $\gamma \in (0, 1)$ where $A \in \mathbb{R}^{m \times n}$ is a known sampling matrix and \mathbf{W} is a standard Gaussian random vector. For any distortion $\alpha \in [0, 1]$, the sparsity pattern estimate $\hat{\mathbf{S}}$ that minimizes the error probability $\Pr\{d(\mathbf{S}, \hat{\mathbf{S}}) > \alpha\}$ corresponds to the estimator

$$\hat{\mathbf{S}}_{\gamma, \alpha}^{\text{OPT}}(\mathbf{y}, A, k) \in \arg \max_{\mathbf{s} \in \mathcal{S}_k^n} \sum_{\mathbf{s}' \in B_\alpha(\mathbf{s})} \exp\{\psi(\mathbf{s}')\} \quad (28)$$

where $B_\alpha(\mathbf{s}) = \{\mathbf{s}' \in \mathcal{S}_k^n : d(\mathbf{s}, \mathbf{s}') \leq \alpha\}$ and

$$\psi(\mathbf{s}) = -\frac{1}{2} \left[\|\Sigma_{\mathbf{s}}^{-1/2} \mathbf{y}\|^2 + \log |\Sigma_{\mathbf{s}}| \right]$$

with $\Sigma_{\mathbf{s}} = \gamma A_{\mathbf{s}} A_{\mathbf{s}}^T + (1-\gamma) I_m$.

Unlike the nearest subspace and thresholding estimators, the optimal Gaussian estimator (28) depends on the SNR parameter γ as well as the distortion bound α . One consequence of this dependence is that the optimal estimate is not invariant to scalings of the samples \mathbf{y} . Another consequence is that it is not possible in general for a single sparsity pattern estimate $\hat{\mathbf{s}}$ to be uniformly optimal for all distortions $\alpha \in [0, 1]$.

To determine whether or not the optimal Gaussian estimator is significantly better than estimators that do not depend on the parameters γ or α , it is useful to consider the bounds on the operational sampling rate distortion functions illustrated in Figures 2 and 3. Since the lower bounds in these figures apply universally to any estimator, including the optimal Gaussian estimator, the relative tightness of the upper and lower bounds shows that near-optimal performance can be attained using the nearest subspace estimator in the high SNR setting and the thresholding estimator in the low SNR setting.

One explanation for this behavior is provided by the following non-asymptotic result which shows that the optimal Gaussian estimator converges pointwise to the nearest subspace estimator as $\gamma \rightarrow 1$ and to the thresholding estimator as $\gamma \rightarrow 0$. The proof of this result is rather long and can be found in [43].

Theorem 5. *Let $\mathbf{y} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ and $k \in \{1, 2, \dots, n\}$ be fixed inputs. Let $\alpha \in [0, 1)$ be a fixed distortion.*

1) *If the nearest subspace estimator is unique, then*

$$\lim_{\gamma \rightarrow 1} \hat{\mathbf{s}}_{\gamma, \alpha}^{\text{OPT}}(\mathbf{y}, A, k) = \hat{\mathbf{s}}^{\text{NS}}(\mathbf{y}, A, k). \quad (29)$$

2) *If the thresholding estimator is unique and the columns of A are equal magnitude, then*

$$\lim_{\gamma \rightarrow 0} \hat{\mathbf{s}}_{\gamma, \alpha}^{\text{OPT}}(\mathbf{y}, A, k) = \hat{\mathbf{s}}^{\text{TH}}(\mathbf{y}, A, k). \quad (30)$$

We first address the high SNR convergence (29) shown in Theorem 5. For the special case of exact recovery ($\alpha = 0$), this result is not particularly surprising, in part because the nearest subspace estimator corresponds to the maximum likelihood estimate of the vector \mathbf{x} . What is less obvious, and in fact significantly more difficult to prove, is that this convergence occurs universally for all distortion bounds α . This result is interesting because it explains why the nearest subspace algorithm performs so well in the high SNR setting for a large range of distortions. Nevertheless, it is important to keep in mind that the rate at which the probability of error tends to zero still depends on the distortion bound α (see for instance, the example given in [43, Proposition 2]).

Next, we address the low SNR convergence (30) shown in Theorem 5. In this case, the log likelihood function $\exp\{\psi(\mathbf{s})\}$ becomes proportional to $\|\mathbf{A}_\mathbf{s} \mathbf{y}\|^2$ as $\gamma \rightarrow 0$. Using the fact that $\|\mathbf{A}_\mathbf{s} \mathbf{y}\|^2$ can be expressed as a sum over terms indexed by the indices in \mathbf{s} shows that the convergence occurs uniformly for all α . What is particularly interesting about this result is that it shows that near-optimal estimation in the low SNR setting can be achieved using a computationally efficient estimator.

APPENDIX A PROOF OF THEOREM 1

Let \mathbf{s}^* denote the true sparsity pattern of \mathbf{x} and define the set $B_\alpha := \{\mathbf{s} \in \mathcal{S}_k^n : d(\mathbf{s}^*, \mathbf{s}) \leq \alpha\}$. Observe that the event

$$\mathcal{E} := \left\{ \min_{\mathbf{s} \in B_\alpha} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 < \min_{\mathbf{s} \in B_\alpha^c} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 \right\}$$

guarantees that the distortion of the nearest subspace estimate is less than or equal to α . Moreover, for any threshold t , the event \mathcal{E} is implied by $\mathcal{E}_1(t) \cap \mathcal{E}_2(t)$ where

$$\begin{aligned} \mathcal{E}_1(t) &= \left\{ \min_{\mathbf{s} \in B_\alpha} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 < (m-k) \cdot t \right\}, \\ \mathcal{E}_2(t) &= \left\{ \min_{\mathbf{s} \in B_\alpha^c} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 > (m-k) \cdot t \right\}, \end{aligned}$$

and thus the probability of error can be upper bounded as

$$P_e^{(n)} \leq \Pr\{\mathcal{E}^c\} \leq \Pr\{\mathcal{E}_1^c(t)\} + \Pr\{\mathcal{E}_2^c(t)\}.$$

In the following, we show that there exists a threshold t such that both $\Pr\{\mathcal{E}_1^c(t)\}$ and $\Pr\{\mathcal{E}_2^c(t)\}$ decay exponentially rapidly with the problem size n when the elements of the sampling matrix \mathbf{A} are i.i.d. zero-mean Gaussian random variables.

We begin with following result which characterizes the marginal distribution of each projection $\|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2$. For two sets \mathbf{s} and \mathbf{u} we use $\mathbf{s} \setminus \mathbf{u}$ to denote the difference set $\{s \in \mathbf{s} : s \notin \mathbf{u}\}$. Additionally, we use χ_d^2 to denote a chi-square random variable with d degrees of freedom.

Lemma 2. *For any sparsity pattern $\mathbf{s} \in \mathcal{S}_k^n$, the random variable*

$$\left(1 + \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}}\|^2\right)^{-1} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2$$

has a chi-square distribution with $m - k$ degrees of freedom.

Proof: Since $\Pi(\mathbf{A}_\mathbf{s})$ projects onto the null space of $\mathbf{A}_\mathbf{s}$,

$$\begin{aligned} \Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y} &= \Pi(\mathbf{A}_\mathbf{s}) [\mathbf{A}_{\mathbf{s}^*} \mathbf{x}_{\mathbf{s}^*} + \mathbf{W}] \\ &= \Pi(\mathbf{A}_\mathbf{s}) [\mathbf{A}_{\mathbf{s}^* \setminus \mathbf{s}} \mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}} + \mathbf{W}] \end{aligned}$$

where the vector $\mathbf{A}_{\mathbf{s}^* \setminus \mathbf{s}} \mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}} + \mathbf{W}$ is independent of $\Pi(\mathbf{A}_\mathbf{s})$ and has i.i.d. Gaussian elements with zero mean and variance $1 + \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}}\|^2$. Using the rotational invariance of the Gaussian distribution, and the fact that the projection matrix $\Pi(\mathbf{A}_\mathbf{s})$ has rank $m - k$ almost surely concludes the proof. ■

We now consider the event $\mathcal{E}_1(t)$. Using Lemma 2 gives

$$\begin{aligned} \Pr\{\mathcal{E}_1^c(t)\} &= \Pr\left\{ \min_{\mathbf{s} \in B_\alpha} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 \geq (m-k) \cdot t \right\} \\ &\leq \Pr\left\{ \|\Pi(\mathbf{A}_{\mathbf{s}^*}) \mathbf{Y}\|^2 \geq (m-k) \cdot t \right\} \\ &= \Pr\left\{ \chi_{m-k}^2 \geq (m-k) \cdot t \right\}. \end{aligned}$$

Applying the chi-square concentration bounds in Lemma 8 in Appendix F shows that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr\{\mathcal{E}_1^c(t)\} < 0 \quad \text{for any } t > 1. \quad (31)$$

Next we consider the event $\mathcal{E}_2(t)$. Using the union bound and Lemma 2 gives

$$\begin{aligned} \Pr\{\mathcal{E}_2^c(t)\} &= \Pr\left\{ \min_{\mathbf{s} \in B_\alpha^c} \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 \leq (m-k) \cdot t \right\} \\ &\leq |B_\alpha^c| \max_{\mathbf{s} \in B_\alpha^c} \Pr\left\{ \|\Pi(\mathbf{A}_\mathbf{s}) \mathbf{Y}\|^2 \leq (m-k) \cdot t \right\} \\ &= |B_\alpha^c| \max_{\mathbf{s} \in B_\alpha^c} \Pr\left\{ \chi_{m-k}^2 \leq \frac{(m-k) \cdot t}{1 + \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}}\|^2} \right\} \end{aligned}$$

If we defined the set $U_\beta := \{\mathbf{s} : d(\mathbf{s}^*, \mathbf{s}) = \lceil \beta k \rceil / k\}$ and the functional

$$P_\beta(\mathbf{x}) := \min_{\mathbf{s} \in U_\beta} \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}}\|^2, \quad (32)$$

then we obtain the further bound

$$\begin{aligned} \Pr\{\mathcal{E}_2^c(t)\} &\leq k \max_{\alpha \leq \beta \leq 1} |U_\beta| \Pr\left\{ \chi_{m-k}^2 \leq \frac{(m-k) \cdot t}{1 + \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^* \setminus \mathbf{s}}\|^2} \right\} \\ &= k \max_{\alpha \leq \beta \leq 1} |U_\beta| \Pr\left\{ \chi_{m-k}^2 \leq \frac{(m-k) \cdot t}{1 + P_\beta(\mathbf{x})} \right\}. \end{aligned} \quad (33)$$

A simple counting argument gives

$$|U_\beta| = \binom{k}{\lceil \beta k \rceil} \binom{n-k}{\lceil \beta k \rceil},$$

and using Lemma 7 in Appendix F shows that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{U}_\beta| = \Omega H(\beta) - (1-\Omega)H\left(\frac{\Omega\beta}{1-\Omega}\right). \quad (34)$$

To lower bound $P_\beta(\mathbf{x})$ we use the following result.

Lemma 3. *For any sequence of vectors $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$ and $\epsilon \in (0, 1)$,*

$$\liminf_{n \rightarrow \infty} P_\beta(\mathbf{x}^{(n)}) \geq P(\Omega, F_{\beta-\epsilon})$$

uniformly for all $\beta \in (\epsilon, 1)$ where $P(\cdot, \cdot)$ and F_β are defined by (5) and (6) respectively.

Proof: For each n , let F_n and \tilde{F}_n denote the empirical distributions of $\{x_i : i \in \mathbf{s}^*\}$ and $\{x_{\tilde{i}}^2 : i \in \mathbf{s}^*\}$ respectively. Furthermore, let $\tilde{F}_n^{-1}(p) = \inf\{x : \tilde{F}_n(x) \leq p\}$ denote the inverse of \tilde{F}_n and observe that

$$P_\beta(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{\lceil \beta k \rceil} \tilde{F}_n^{-1}(i/\lceil \beta k \rceil) \geq \frac{k}{n} \cdot \int_0^\beta \tilde{F}_n^{-1}(p) dp. \quad (35)$$

Next, let X be a random variable with distribution F and let \tilde{F} be the distribution of X^2 . By definition, $F_n \rightarrow F$ uniformly for any sequence drawn from the vector source $\mathcal{X}(\Omega, F)$. We note that in some cases, this convergence is sufficient to show that $\tilde{F}_n^{-1} \rightarrow \tilde{F}^{-1}$ uniformly. In general, however, flat sections in F may correspond to discontinuities in \tilde{F}^{-1} and thus more care is needed.

Observe that the convergence $F_n \rightarrow F$ does guarantee that for n large enough,

$$\tilde{F}_n^{-1}(p) \geq \begin{cases} \tilde{F}^{-1}(p - \epsilon), & \text{if } p \geq \epsilon \\ 0, & \text{if } p < \epsilon \end{cases} \quad (36)$$

for all $0 \leq p \leq 1 - \epsilon$. Hence, combining (35) and (36) gives,

$$\liminf_{n \rightarrow \infty} P_\beta(\mathbf{x}^{(n)}) \geq \Omega \int_\epsilon^\beta \tilde{F}^{-1}(p - \epsilon) dp = P(\Omega, F_{\beta-\epsilon})$$

which completes the proof. \blacksquare

Applying Lemma 3 and the chi-square large deviation bounds from Lemma 8 in Appendix F shows that for any $\epsilon > 0$ and $t < 1 + P(\Omega, F_{\beta-\epsilon})$,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr \left\{ \chi_{m-k}^2 < \frac{(m-k) \cdot t}{1 + P_\beta(\mathbf{x}^{(n)})} \right\} \\ \leq (\rho - \Omega) \mathcal{L} \left(\frac{1 + P(\Omega, F_{\beta-\epsilon})}{t} \right). \end{aligned} \quad (37)$$

Combining (33), (34) and (37) shows that the probability of error decays exponentially rapidly with n provided that the sampling rate ρ obeys

$$\rho > \Omega + \max_{\beta \in [\alpha, 1]} \frac{\Omega H(\beta) - (1-\Omega)H\left(\frac{\Omega\beta}{1-\Omega}\right)}{\mathcal{L} \left(\frac{1 + P(\Omega, F_{\beta-\epsilon})}{t} \right)}$$

for some $\epsilon > 0$ and threshold $1 < t < 1 + P(\Omega, F_{\beta-\epsilon})$. Taking the infimum over all such rates completes the proof.

APPENDIX B PROOF OF THEOREM 2

For each integer n , denote by $\mathbf{Z}_n = [Z_{n,1}, \dots, Z_{n,n}]$ the random vector with elements given by

$$Z_{n,i} = \sqrt{\frac{n}{m(1+\|\mathbf{x}\|^2/n)}} \langle \mathbf{A}_i, \mathbf{Y} \rangle,$$

and define the functions

$$\begin{aligned} D_n^-(t) &= \frac{1}{n-k} \sum_{i \notin \mathbf{s}} \mathbf{1}(Z_{n,i}^2 > t) \\ D_n^+(t) &= \frac{1}{k} \sum_{i \in \mathbf{s}} \mathbf{1}(Z_{n,i}^2 < t). \end{aligned}$$

With a bit of work it can be verified that the thresholding estimate is successful with respect to distortion α if and only if

$$\inf_t \max \left\{ \frac{n-k}{k} D_n^-(t), D_n^+(t) \right\} \leq \alpha. \quad (38)$$

In the following, we identify the infimum over all sampling rates ρ such that (38) occurs with probability tending to one as $n \rightarrow \infty$ when the elements of the sampling matrix \mathbf{A} are i.i.d. zero-mean Gaussian random variables.

The key technical parts of the proof are given by the following results which characterize the asymptotic convergence of the empirical distributions $D_n^+(t)$ and $D_n^-(t)$. The proofs are given in Sections B-A and B-B below.

Lemma 4. *For any sequence $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$ and $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} \Pr \left\{ \sup_t |D_n^-(t) - D^+(t)| > \epsilon \right\} = 0$$

where $D^-(t) = 2Q(t)$.

Lemma 5. *For any sequence $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$ and $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} \Pr \left\{ \sup_t |D_n^+(t) - D^+(t; \Omega, F, \rho)| > \epsilon \right\} = 0$$

where $D^+(t; \Omega, F, \rho) = \int_{\mathbb{R}} G\left(\frac{\rho u^2}{1+P(\Omega, F)}, t\right) dF(u)$.

Using Lemmas 4 and 5, the remainder of the proof is relatively straightforward. Since $D^-(t)$ is continuous and strictly decreasing, and since $D^+(t; \Omega, F, \rho)$ is continuous and strictly increasing in t , and continuous and strictly decreasing in ρ , the bound defined in Theorem 2 can be rewritten as

$$\begin{aligned} \rho^{\text{TH-UB}}(\alpha) &= \inf \left\{ \rho : D^+(t^*; \Omega, F, \rho) < \alpha \right\} \quad \text{where} \\ t^* &= \inf \left\{ t : \frac{(1-\Omega)}{\Omega} D^-(t) < \alpha \right\}. \end{aligned}$$

The convergence given in Lemmas 4 and 5 shows that condition (38) occurs with probability tending to one if $\rho > \rho^{\text{TH-UB}}(\alpha)$ and does not occur with probability tending to one if $\rho < \rho^{\text{TH-UB}}(\alpha)$, which concludes the proof.

A. Proof of Lemma 4

Conditioned on any realization $\mathbf{Y}^{(n)} = \mathbf{y}^{(n)}$, the variables $\{Z_{n,i} : i \notin \mathbf{s}\}$ are i.i.d. Gaussian with zero mean and variance

$\sigma^2(\mathbf{y}^{(n)}, \mathbf{x}^{(n)}) = (1 + \frac{1}{n} \|\mathbf{x}^{(n)}\|^2)^{-1} \|\mathbf{y}^{(n)}\|^2$. By the Glivenko-Cantelli theorem,

$$D_n^-(t) \rightarrow \Pr \left\{ \chi_1^2 > \frac{t}{\sigma^2(\mathbf{y}^{(n)}, \mathbf{x}^{(n)})} \right\}$$

almost surely and uniformly as $n \rightarrow \infty$.

Furthermore, by law of large numbers, $\sigma^2(\mathbf{Y}^{(n)}, \mathbf{x}^{(n)}) \rightarrow 1$ almost surely as $n \rightarrow \infty$ for any random sequence $\{\mathbf{Y}^{(n)}\}$, and thus,

$$\Pr \left\{ \chi_1^2 > \frac{t}{\sigma^2(\mathbf{Y}^{(n)}, \mathbf{x}^{(n)})} \right\} \rightarrow \Pr\{\chi_1^2 > t\}$$

almost surely and uniformly as $n \rightarrow \infty$. Using the fact that $\Pr\{\chi_1^2 > t\} = 2Q(t)$ for $t \geq 0$ completes the proof.

B. Proof of Lemma 5

The main challenge in this proof is that, for each problem of size n , the variables $\{Z_{n,i}^2 : i \in \mathbf{s}\}$ are neither independent nor identically distributed. To proceed, we first show that $D_n^+(t)$ converges in expectation to the limit $D^+(t)$, and then show that $D_n^+(t)$ also converges in probability.

Convergence in expectation: Observe that the expectation of $D_n^+(t)$ depends only on the marginal distributions of the elements $\{Z_{n,i}^2 : i \in \mathbf{s}\}$ and is given by

$$\mathbb{E}[D_n^+(t)] = \frac{1}{k} \sum_{i \in \mathbf{s}} \Pr\{Z_{n,i}^2 < t\}.$$

To further characterize the probabilities $\Pr\{Z_{n,i}^2 < t\}$, we decompose each column of the sampling matrix as $\mathbf{A}_i = \|\mathbf{A}_i\| \mathbf{U}_i$ where $\mathbf{U}_i = \mathbf{A}_i / \|\mathbf{A}_i\|$ is a random unit vector independent of $\|\mathbf{A}_i\|$. We can then write

$$\begin{aligned} \mathbf{A}_i^T \mathbf{Y} &= \|\mathbf{A}_i\|^2 x_i + \mathbf{A}_i^T (\mathbf{Y} - \mathbf{A}_i x_i) \\ &= \|\mathbf{A}_i\|^2 x_i + \|\mathbf{A}_i\| \mathbf{U}_i^T (\mathbf{Y} - \mathbf{A}_i x_i) \end{aligned} \quad (39)$$

where the random variable $\mathbf{U}_i^T (\mathbf{Y} - \mathbf{A}_i x_i)$ is independent of $\|\mathbf{A}_i\|$ and has a Gaussian distribution with zero mean and variance $1 + \|\mathbf{x}\|^2/n - x_i^2/n$.

Since $n\|\mathbf{A}_i\|^2$ is a chi-square random variable with m degrees of freedom, the chi-square large deviation results in Lemma 8 in Appendix F show that

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} \left| \|\mathbf{A}_i^{(n)}\|^2 - \rho \right| = 0 \quad (40)$$

almost surely for any sequence of sampling matrices $\mathbf{A}^{(n)}$.

Combining (39) and (40) shows that the marginal distributions of the elements $\{Z_{n,i} : i \in \mathbf{s}\}$ are asymptotically Gaussian, or more explicitly,

$$\lim_{n \rightarrow \infty} \max_{i \in \mathbf{s}} \sup_t \left| \Pr(Z_{n,i} > t) - Q\left(\frac{t - \mu_{n,i}}{\sigma_{n,i}}\right) \right| = 0$$

where the mean and variance are given by

$$\mu_{n,i} = \sqrt{\frac{\rho}{1 + \|\mathbf{x}\|^2/n}} x_i \quad \text{and} \quad \sigma_{n,i}^2 = \frac{1 + \|\mathbf{x}\|^2/n - x_i^2/n}{1 + \|\mathbf{x}\|^2/n}.$$

Hence, by the definition of $G(\cdot, \cdot)$, the asymptotic expectation of $D_n^+(t)$ can be expressed as

$$\lim_{n \rightarrow \infty} \mathbb{E}[D_n^+(t)] = \lim_{n \rightarrow \infty} \frac{1}{k} \sum_{i \in \mathbf{s}} G\left(\frac{\mu_{n,i}^2}{\sigma_{n,i}^2}, \frac{t}{\sigma_{n,i}^2}\right) \quad (41)$$

To evaluate the right hand side of (41), we use a truncation argument to show that the effect of any “large” nonzero elements x_i is negligible. More precisely, we define the set of indices $\mathcal{I}_n = \{i \in \mathbf{s} : x_i^2 < \sqrt{n}\}$ and observe that since the empirical distribution F_n of $\{x_i : i \in \mathbf{s}\}$ converges uniformly to F for any sequence $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \max_{i \in \mathcal{I}_n} |\sigma_{n,i}^2 - 1| &= 0 \\ \lim_{n \rightarrow \infty} \max_{i \in \mathcal{I}_n} |\mu_{n,i} - \sqrt{\frac{\rho}{1 + P(\Omega, F)}} x_i| &= 0, \end{aligned}$$

and

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{k} \sum_{i \notin \mathcal{I}_n} G\left(\frac{\mu_{n,i}^2}{\sigma_{n,i}^2}, \frac{t}{\sigma_{n,i}^2}\right) \\ \leq \lim_{n \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k \mathbf{1}(x_i^2 > \sqrt{n}) \\ = \lim_{n \rightarrow \infty} 1 - F(n^{1/4}) + F(-n^{1/4}) = 0 \end{aligned}$$

where we use the fact that $0 \leq G(\cdot, \cdot) \leq 1$.

Starting from Equation (41) and using the above observations gives

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}[D_n^+(t)] &= \lim_{n \rightarrow \infty} \frac{1}{k} \sum_{i \in \mathcal{I}_n} G\left(\frac{\rho}{1 + P(\Omega, F)} x_i^2, t\right) \\ &= \lim_{n \rightarrow \infty} \int_0^{\sqrt{n}} G\left(\frac{\rho}{1 + P(\Omega, F)} u^2, t\right) dF_n(u) \\ &= \lim_{n \rightarrow \infty} \int_0^{\sqrt{n}} G\left(\frac{\rho}{1 + P(\Omega, F)} u^2, t\right) dF(u) \\ &= \int_{\mathbb{R}} G\left(\frac{\rho}{1 + P(\Omega, F)} u^2, t\right) dF(u), \end{aligned}$$

and thus we have shown that $D_n^+(t)$ converges in expectation to the desired limit.

Convergence in probability: We now show that the variance of $D_n^+(t)$ tends to zero uniformly as $n \rightarrow \infty$. Using this result, the desired convergence in probability follows directly from Chebyshev’s inequality.

To begin, we define the set $\mathcal{I}_n = \{i \in \mathbf{s} : x_i^2 < \sqrt{n}\}$ and the quantity

$$\begin{aligned} \delta_{i,j}^{(n)}(t) &:= \left| \Pr(Z_{n,i}^2 < t, Z_{n,j}^2 < t) \right. \\ &\quad \left. - \Pr(Z_{n,i}^2 < t) \Pr(Z_{n,j}^2 < t) \right|. \end{aligned}$$

Since $|\mathcal{I}_n|/k \rightarrow 1$ for any sequence $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$, and since $0 \leq \delta_{i,j}^{(n)} \leq 1$, we can bound the variance of $D_n^+(t)$ as

$$\begin{aligned} \limsup_{n \rightarrow \infty} \text{Var}(D_n^+(t)) &\leq \limsup_{n \rightarrow \infty} \frac{1}{k^2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \delta_{i,j}^{(n)}(t) \\ &= \limsup_{n \rightarrow \infty} \frac{1}{k^2} \sum_{i,j \in \mathcal{I}_n : i \neq j} \delta_{i,j}^{(n)}(t) \\ &\leq \limsup_{n \rightarrow \infty} \max_{i,j \in \mathcal{I}_n : i \neq j} \delta_{i,j}^{(n)}(t). \end{aligned} \quad (42)$$

The next step is to show that the right hand side of (42) converges to zero uniformly in t . For a given n , let p_i denote

the probability measure on $Z_{n,i}$ and $p_{i,j}$ the joint probability measure on $(Z_{n,i}, Z_{n,j})$. Then, for all t ,

$$\delta_{i,j}^{(n)}(t) \leq \|p_{i,j} - p_i p_j\|_{TV}$$

where $\|p_{i,j} - p_i p_j\|_{TV}$ denotes the total variation distance between $p_{i,j}$ and $p_i p_j$. Using Pinsker's inequality [44], the total variation distance can be upper bounded by the mutual information between $Z_{n,i}$ and $Z_{n,j}$ as

$$\|p_{i,j} - p_i p_j\|_{TV}^2 \leq 2I(Z_{n,i}; Z_{n,j}).$$

To bound the mutual information $I(Z_{n,i}; Z_{n,j})$ we may expand the information $I(Z_{n,i}, \mathbf{A}_i; Z_{n,j}, \mathbf{A}_j)$ two different ways using the chain rule for mutual information to obtain

$$\begin{aligned} I(Z_{n,i}; Z_{n,j}) &\leq I(Z_{n,i}; Z_{n,j} | \mathbf{A}_i, \mathbf{A}_j) + I(Z_{n,i}; \mathbf{A}_j | \mathbf{A}_i) \\ &\quad + I(\mathbf{A}_i; Z_{n,j} | \mathbf{A}_j) + I(\mathbf{A}_i; \mathbf{A}_j). \end{aligned}$$

In the following, we show that each of the terms on the right hand side of the above equation tends to zero as $n \rightarrow \infty$.

First, by the independence of the columns \mathbf{A}_i and \mathbf{A}_j we have $I(\mathbf{A}_i; \mathbf{A}_j) = 0$ for all n .

Next, we consider the term $I(Z_{n,i}; \mathbf{A}_j | \mathbf{A}_i)$. Observe that the random variable $\mathbf{A}_i^T \mathbf{Y}$ can be decomposed as

$$\mathbf{A}_i^T \mathbf{Y} = \mathbf{A}_i^T (\mathbf{Y} - \mathbf{A}_j x_j) + \mathbf{A}_i^T \mathbf{A}_j x_j$$

where $\mathbf{A}_i^T (\mathbf{Y} - \mathbf{A}_j x_j)$ is independent of \mathbf{A}_j . Therefore, since $Z_{n,i}$ is proportional to $\mathbf{A}_i^T \mathbf{Y}$, we can write

$$I(Z_{n,i}; \mathbf{A}_j | \mathbf{A}_i) = I(\mathbf{A}_i^T \mathbf{Y}; \mathbf{A}_i^T \mathbf{A}_j x_j | \mathbf{A}_i).$$

Conditioned on any realization $\mathbf{A}_i = \mathbf{a}_i$, the variables $\mathbf{a}_i^T \mathbf{Y}$ and $\mathbf{a}_i^T \mathbf{A}_j x_j$ are jointly Gaussian with covariance

$$\Sigma = \|\mathbf{a}_i\|^2 \begin{bmatrix} 1 + \|\mathbf{x}\|^2/n - x_i^2/n & x_j^2/n \\ x_j^2/n & x_j^2/n \end{bmatrix}.$$

From the definition of \mathcal{I}_n , we conclude that

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \max_{i,j \in \mathcal{I}_n: i \neq j} I(Z_{n,i}; \mathbf{A}_j | \mathbf{A}_i) \\ &= \limsup_{n \rightarrow \infty} \max_{i,j \in \mathcal{I}_n: i \neq j} \log \left(1 + \frac{x_j^2/n}{1 + \|\mathbf{x}\|^2/n - (x_i^2 + x_j^2)/n} \right) \\ &= 0. \end{aligned}$$

A symmetric argument shows that the same limit applies to the term $I(\mathbf{A}_i; Z_{n,j} | \mathbf{A}_j)$.

Lastly, we consider the term $I(Z_{n,i}; Z_{n,j} | \mathbf{A}_i, \mathbf{A}_j)$. With a bit of work, it can be shown that for any realizations $\mathbf{A}_i = \mathbf{a}_i, \mathbf{A}_j = \mathbf{a}_j$, the variables $Z_{n,i}$ and $Z_{n,j}$ are jointly Gaussian with covariance

$$\Sigma = \frac{1 + \|\mathbf{x}\|^2/n - (x_i^2 + x_j^2)/n}{(m/n)(1 + \|\mathbf{x}\|^2/n)} \begin{bmatrix} \|\mathbf{a}_i\|^2 & \mathbf{a}_i^T \mathbf{a}_j \\ \mathbf{a}_i^T \mathbf{a}_j & \|\mathbf{a}_j\|^2 \end{bmatrix}.$$

Using the fact that $U = \left(\frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \right)^2$ has a Beta(1, $n-1$) distribution, it follows that

$$\begin{aligned} I(Z_{n,i}; Z_{n,j} | \mathbf{A}_i, \mathbf{A}_j) &= \frac{1}{2} \mathbb{E} \log \left(\frac{1}{1-U} \right) \\ &\leq \frac{1}{2} \mathbb{E} \frac{U}{1-U} \\ &= \frac{1}{2(n-2)}. \end{aligned}$$

Hence,

$$\limsup_{n \rightarrow \infty} \max_{i,j \in \mathcal{I}_n: i \neq j} I(Z_{n,i}; Z_{n,j} | \mathbf{A}_i, \mathbf{A}_j) = 0$$

which completes the proof.

C. Proof of Corollary 2

From Theorem 2, a sampling rate distortion pair (ρ, α) is achievable if

$$\Pr \left\{ |W + \sqrt{\frac{\rho}{1+P(\Omega, F)}} X| < t \right\} \leq \alpha \quad (43)$$

where W has a Gaussian distribution with zero mean and unit variance, X has distribution F and is independent of W , and $t = Q^{-1}(\frac{\alpha\Omega}{2(1-\Omega)})$. Observe that for any $x > 0$, we can write

$$\begin{aligned} &\Pr \left\{ |W + \sqrt{\frac{\rho}{1+P(\Omega, F)}} X| < t \right\} \\ &\leq \Pr\{|X| > x\} + \Pr \left\{ |W + \sqrt{\frac{\rho}{1+P(\Omega, F)}} X| < t \mid |X| \leq x \right\}. \end{aligned}$$

Furthermore, if we let $x = \sqrt{P(1, F_{\alpha/2})}$, then

$$x^2 = (2/\alpha) \int_0^{\alpha/2} F_{X^2}^{-1}(p) dp \leq F_{X^2}^{-1}(\alpha/2)$$

where $F_{X^2}^{-1}$ denotes the quantile function of X^2 , and thus

$$\Pr\{|X| > x\} \leq \alpha/2. \quad (44)$$

Additionally, if ρ satisfies (12), then

$$\begin{aligned} &\Pr \left\{ |W + \sqrt{\frac{\rho}{1+P(\Omega, F)}} X| < t \mid |X| \leq x \right\} \\ &\leq \Pr \left\{ W < t - \sqrt{\frac{\rho}{1+P(\Omega, F)}} x \right\} \\ &= Q \left(\sqrt{\frac{\rho}{1+P(\Omega, F)}} x - t \right) \\ &\leq \alpha/2. \end{aligned} \quad (45)$$

Combining (44) and (45) shows that the left hand side of (43) is less than or equal to α which completes the proof.

APPENDIX C PROOF OF THEOREM 3

Let (ρ, α) be a sampling rate distortion pair that is achievable for the source $\mathcal{X}(\Omega, F)$ and estimator $\hat{\mathbf{s}}$ and let $\{\mathbf{A}^{(n)} \in \mathbb{R}^{\lceil \rho \cdot n \rceil \times n}\}$ be the sampling matrix sequence that achieves this rate. Furthermore, let $\{\mathbf{x}^{(n)}\} \in \mathcal{X}(\Omega, F)$ be an arbitrary sequence of vectors with sparsity k_n , and let \tilde{k}_n be a sequence of integers that obeys $\lim_{n \rightarrow \infty} |k_n - \tilde{k}_n|/n = 0$ and $\limsup_{n \rightarrow \infty} k_n - \tilde{k}_n \leq 0$. In the following, we show that the asymptotic performance of the estimator $\hat{\mathbf{s}}(\mathbf{y}, \mathbf{A}, \tilde{k}_n)$ corresponding to the sequence \tilde{k}_n is equal to that of the estimator $\hat{\mathbf{s}}(\mathbf{y}, \mathbf{A}, k_n)$ corresponding to the true sparsity k_n . For notational simplicity, we will often make the dependence on the problem size n implicit and write \mathbf{x} , \mathbf{A} , and k instead of $\mathbf{x}^{(n)}$, $\mathbf{A}^{(n)}$, and k_n .

To begin, let $\{\tilde{\mathbf{x}}^{(n)}\} \in \mathcal{X}(\Omega, F)$ be a related sequence of vectors with sparsity \tilde{k} whose sparsity pattern $\tilde{\mathbf{s}}$ obeys $|\tilde{\mathbf{s}} \cap \mathbf{s}| = \min(k, \tilde{k})$ and whose nonzero values obey $\|\tilde{\mathbf{x}} - \mathbf{x}\| \rightarrow 0$ as $n \rightarrow \infty$. By the definition of $\mathcal{X}(\Omega, F)$, it can be verified that such a sequence is guaranteed to exist. Also, for each integer

n , let $\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W}$ and $\tilde{\mathbf{Y}} = \mathbf{A}\tilde{\mathbf{x}} + \mathbf{W}$ be the samples of \mathbf{x} and $\tilde{\mathbf{x}}$ corresponding to the common sampling matrix \mathbf{A} and noise vector \mathbf{W} .

Now, using the triangle inequality gives

$$d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \tilde{k})) \leq d(\mathbf{s}, \tilde{\mathbf{s}}) + d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \tilde{k})),$$

where, by assumption, the first term on the right obeys

$$d(\mathbf{s}, \tilde{\mathbf{s}}) = \max(1 - \tilde{k}/k, 1 - k/\tilde{k}) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

To deal with the second term, we note that

$$d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \tilde{k})) = d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}(\tilde{\mathbf{Y}} + \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}}), \mathbf{A}, \tilde{k})), \quad (46)$$

and use the following lemma.

Lemma 6. *Let $\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W}$ where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$ are fixed and $\mathbf{W} \in \mathbb{R}^m$ is a standard Gaussian vector. Furthermore, let $\hat{\mathbf{s}} : \mathbb{R}^m \mapsto \mathcal{S}^n$ be a sparsity pattern estimator. If $\Pr\{d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y})) > \alpha\} < \epsilon$, then*

$$\Pr\{d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y} + \mathbf{z})) \geq \alpha\} < \epsilon + \|\mathbf{z}\| \quad (47)$$

for all $\mathbf{z} \in \mathbb{R}^m$.

Proof: For simplicity, suppose that $\hat{\mathbf{s}}$ is a deterministic function; the extension of the proof to handle random estimators is straightforward. Define the set

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^m : d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{A}\mathbf{x} + \mathbf{w})) > \alpha\},$$

and observe that $\Pr\{\mathbf{W} \in \mathcal{W}\} < \epsilon$. Thus, for any $\mathbf{z} \in \mathbb{R}^m$,

$$\begin{aligned} \Pr\{d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y} + \mathbf{z})) > \alpha\} &= \Pr\{\mathbf{W} + \mathbf{z} \in \mathcal{W}\} \\ &\leq \epsilon + |\Pr\{\mathbf{W} + \mathbf{z} \in \mathcal{W}\} - \Pr\{\mathbf{W} \in \mathcal{W}\}| \\ &\leq \epsilon + \sup_{\mathcal{A} \subseteq \mathbb{R}^m} |\Pr\{\mathbf{W} + \mathbf{z} \in \mathcal{A}\} - \Pr\{\mathbf{W} \in \mathcal{A}\}| \\ &\leq \epsilon + \sqrt{2D_{\text{KL}}(\mathbf{W} + \mathbf{z} \| \mathbf{W})} \\ &= \epsilon + \|\mathbf{z}\| \end{aligned} \quad (48)$$

where (48) follows from Pinsker's inequality (see for example [44]) and $D_{\text{KL}}(\cdot \| \cdot)$ is the Kullback-Leibler divergence. ■

Combining Lemma 6 and (46) gives

$$\begin{aligned} \Pr\{d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \tilde{k})) > \alpha\} \\ < \Pr\{d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}(\tilde{\mathbf{Y}}, \mathbf{A}, \tilde{k})) > \alpha\} + \mathbb{E}[\|\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})\|] \end{aligned} \quad (49)$$

where the expectation is taken with respect to the random matrix \mathbf{A} . By the achievability of (ρ, α) , the first term on the right hand side of (49) tends to zero as $n \rightarrow \infty$. Additionally, since the elements of the vector $\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})$ are i.i.d. Gaussian with zero mean and variance $\|\mathbf{x} - \tilde{\mathbf{x}}\|^2/n$, it follows from Jensen's inequality that

$$\mathbb{E}[\|\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})\|] \leq \sqrt{\frac{m\|\mathbf{x} - \tilde{\mathbf{x}}\|^2}{n}},$$

and thus the second term tends to zero as $n \rightarrow \infty$ by the assumptions on $\tilde{\mathbf{x}}$. Hence, we have shown that for any $\epsilon > 0$,

$$\Pr\{\mathbf{s}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \tilde{k}) > \alpha + \epsilon\} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

which proves our desired result.

APPENDIX D PROOF OF THEOREM 4

To begin, we note that the sampling rate associated with the sequence $\{\mathbf{A}^{(n)}\}$ is given by

$$\lim_{n \rightarrow \infty} \frac{[\rho_1 \cdot \lceil \lambda \cdot n \rceil + [\rho_2 \cdot (n - \lceil \lambda \cdot n \rceil)]]}{n} = \lambda \rho_1 + (1 - \lambda) \rho_2.$$

Also, since

$$\mathbb{E}\|\mathbf{A}^{(n)}\|_F^2 = \mathbb{E}\|\mathbf{A}_1^{(\lceil \lambda \cdot n \rceil)}\|_F^2 + \mathbb{E}\|\mathbf{A}_2^{(n - \lceil \lambda \cdot n \rceil)}\|_F^2,$$

each matrix $\mathbf{A}^{(n)}$ obeys the scaling constraint (3).

Next, we observe that for each integer n , the vector of samples \mathbf{Y} can be expressed as

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^{(\lceil \lambda \cdot n \rceil)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2^{(n - \lceil \lambda \cdot n \rceil)} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix},$$

where $\mathbf{X}_1 \in \mathbb{R}^{\lceil \lambda \cdot n \rceil}$ and $\mathbf{X}_2 \in \mathbb{R}^{n - \lceil \lambda \cdot n \rceil}$ correspond to a random permutation and partition of the elements in $\mathbf{x}^{(n)}$. Note that the sparsity patterns \mathbf{S}_1 and \mathbf{S}_2 of \mathbf{X}_1 and \mathbf{X}_2 are random sets that obey $\mathbb{E}|\mathbf{S}_1| = \Omega[\lambda \cdot n]$ and $\mathbb{E}|\mathbf{S}_2| = \Omega(n - \lceil \lambda \cdot n \rceil)$ respectively. Let $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ be estimates of these sparsity patterns that are made separately using the samples \mathbf{Y}_1 and \mathbf{Y}_2 and expected sizes (since the true sizes are unknown), and let $\hat{\mathbf{S}}$ be the estimate of the original sparsity pattern \mathbf{s} based on the union of $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ mapped back into the original index set of \mathbf{x} .

Since \mathbf{S}_1 and \mathbf{S}_2 correspond to a disjoint partition of \mathbf{s} , it can be verified that

$$d(\mathbf{s}, \hat{\mathbf{S}}) \leq \Lambda \cdot d(\mathbf{S}_1, \hat{\mathbf{S}}_1) + (1 - \Lambda) \cdot d(\mathbf{S}_2, \hat{\mathbf{S}}_2) \quad (50)$$

where

$$\Lambda = \frac{\max(|\mathbf{S}_1|, |\hat{\mathbf{S}}_1|)}{\max(|\mathbf{S}_1|, |\hat{\mathbf{S}}_1|) + \max(|\mathbf{S}_2|, |\hat{\mathbf{S}}_2|)}.$$

Also, since the permutation is independent of $\mathbf{x}^{(n)}$, both $\{\mathbf{X}_1^{(n)}\}$ and $\{\mathbf{X}_2^{(n)}\}$ are elements of the vector source $\mathcal{X}(\Omega, F)$ almost surely. Thus, $\lim_{n \rightarrow \infty} \Lambda = \lambda$ almost surely and, by Theorem 3, the distortions on the right hand side of (50) will be less than α_1 and α_2 , respectively, with probability tending to one. Putting everything together shows the distortion $\lambda \alpha_1 + (1 - \lambda) \alpha_2$ is achievable, which concludes the proof.

APPENDIX E PROOFS OF SCALING BOUNDS

A. Proof of Proposition 1

This proof follows from the upper bound (7) given in Theorem 1. We first consider the denominator. Using Lemma 1, Lemma 9 in Appendix F, and the concavity of the logarithm, shows that there exists some constant $C_1 > 0$ such that

$$\mathcal{L}(1 + P(\beta\Omega, F_\beta)) \geq C_1 \log(1 + \beta^{4L+2} P^2).$$

Next we consider the numerator. Since the binary entropy function $H(\cdot)$ is concave and increasing on the interval $[0, 1/2]$, we can write

$$\Omega H(\beta) + (1 - \Omega) H(\frac{\Omega\beta}{1 - \Omega}) \leq 2H(\beta\Omega) \leq 4\beta\Omega \log(\frac{1}{\beta\Omega}).$$

Thus far, we have shown that

$$\rho(\alpha) \leq \Omega + \frac{4}{C_1} \cdot \max_{\alpha \leq \beta \leq 1} \frac{\beta \Omega \log(\frac{1}{\beta \Omega})}{\log(1 + \beta^{4L+2} P^2)}.$$

To conclude the proof, we use Lemma 10 in Appendix F which shows that it is sufficient to take the maximum over only the endpoints $\alpha = \beta$ and $\alpha = 1$.

B. Proof of Proposition 2

This proof follows directly from the upper bound (12) given in Corollary 1. Using Lemma 1 shows that there exists some constant $C_1 < \infty$ such that

$$\frac{1 + P(\Omega, F)}{P(\Omega, F_{\alpha/2})} \leq C_1 \cdot \alpha^{-2L} \left(\frac{1 + P}{P} \right). \quad (51)$$

Furthermore, using the bound $Q^{-1}(p) \leq \sqrt{-2 \log(2p)}$ for $0 \leq p \leq 1/2$ (see [45, pp. 53]) gives

$$\left[Q^{-1}(\frac{\alpha}{2}) + Q^{-1}(\frac{\alpha \Omega}{2(1-\Omega)}) \right]^2 \leq 8 \log \left(\frac{1 - \Omega}{\alpha \Omega} \right). \quad (52)$$

Combining (51) and (52) completes the proof.

APPENDIX F TECHNICAL LEMMAS

Lemma 7 ([44, pp. 151]). *If $k/n \rightarrow p$ as $n \rightarrow \infty$ for some $0 \leq p \leq 1$ then,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{n}{k} = H(p)$$

where $H(p)$ is the binary entropy function.

Lemma 8. *Let χ_d^2 be chi-square random variable with d degrees of freedom. For any $\epsilon > 0$,*

$$\Pr\{\chi_d^2 > (1 + \epsilon)d\} \leq \exp(-d \cdot \epsilon^2/4), \quad (53)$$

$$\Pr\left\{\chi_d^2 < \left(\frac{1}{1+\epsilon}\right)d\right\} \leq \exp(-d \cdot \mathcal{L}(1 + \epsilon)) \quad (54)$$

where $\mathcal{L}(\cdot)$ is defined in (8).

Proof: The proof of (53) follows directly from [46, pp. 1325]. To prove (54), we apply a Chernoff bound (see [44, pp. 318]) to obtain

$$\Pr\left\{\chi_d^2 < \left(\frac{1}{1+\epsilon}\right)d\right\} \leq \exp\left(\mu\left(\frac{1}{1+\epsilon}\right)d\right) \mathbb{E}\left[\exp(-\mu \chi_d^2)\right] \\ = \exp(-d \cdot \Lambda(\mu)).$$

for any $\mu > 0$ where $\Lambda(\mu) = \frac{1}{2} \log(1 + 2\mu) - \mu(\frac{1}{1+\epsilon})$. By differentiating, it can be shown that the maximum of $\Lambda(\mu)$ is attained as $\mu = \epsilon/2$. Noting that $\Lambda(\epsilon/2) = \mathcal{L}(1 + \epsilon)$ completes the proof. ■

Lemma 9. *For any $x \geq 0$,*

$$\mathcal{L}(1 + x) \geq \frac{1}{4} \log(1 + \frac{1}{8}x^2),$$

where $\mathcal{L}(\cdot)$ is defined in (8).

Proof: It is convenient to prove the slightly stronger result

$$\mathcal{L}(1 + x) \geq \frac{1}{4} \log(1 + e^{-2}x^2).$$

Observe that for any $a > 0$, the above expression can be expressed equivalently as

$$-\log\left(\frac{e^a(1 + e^{-2}x^2)}{(1 + x)^2}\right) + a - \frac{2x}{1 + x} > 0.$$

Using the bound $\log(1 + x) < x$, it can be shown that the above statement is true if

$$(1 + a - e^a) + 2ax - (1 - a + e^{a-2})x^2 > 0.$$

Evaluating the above condition with $a = 2$ shows that the bound holds for any $x > 5$ and evaluating with $a = 1$ shows that the bound holds for any $.4 \leq x \leq 5$.

For the case $x < .4$ we use a different bounding technique. Using the bounds $x(1 - x/2) \leq \log(1 + x) \leq x$ it can be shown that

$$\mathcal{L}(1 + x) \geq \frac{x^2(1 - x)}{2(1 + x)} \geq \frac{1}{2}e^{-2}x \geq \frac{1}{2} \log(1 + e^{-2}x^2).$$

Hence, we have shown that the bound holds for all $x \geq 0$. ■

Lemma 10. *Given any $0 < \gamma < \infty$ and $1 \leq b < \infty$, let*

$$\theta(x) = \frac{-x \log(x)}{\log(1 + \gamma x^b)}.$$

Then, for any $0 < \alpha \leq 1/8$,

$$\max_{\alpha \leq x \leq 1/8} \theta(x) < 4 \max\{\theta(\alpha), \theta(1/8)\}$$

Proof: Let $x^* = (8/\gamma)^{1/b}$, $x_1 = \min\{x^*, 1/8\}$, and $x_2 = \max\{\alpha, x^*\}$. Then, observe that

$$\max_{\alpha \leq x \leq 1/8} \theta(x) = \max\left\{\max_{\alpha \leq x \leq x_1} \theta(x), \max_{x_2 \leq x \leq 1/8} \theta(x)\right\}$$

Furthermore,

$$\begin{aligned} & \max_{\alpha \leq x \leq x_1} \theta(x) \\ &= \max_{\alpha \leq x \leq x_1} \left(\frac{-x \log(x)}{\gamma x^b} \cdot \frac{\gamma x^b}{\log(1 + \gamma x^b)} \right) \\ &\leq \left(\max_{\alpha \leq x \leq x_1} \frac{-x \log(x)}{\gamma x^b} \right) \left(\max_{\alpha \leq x \leq x_1} \frac{\gamma x^b}{\log(1 + \gamma x^b)} \right) \\ &= \left(\theta(\alpha) \frac{\log(1 + \gamma \alpha^b)}{\gamma \alpha^b} \right) \left(\frac{\gamma x_1^b}{\log(1 + \gamma x_1^b)} \right) \\ &< 4\theta(\alpha). \end{aligned}$$

Also,

$$\begin{aligned} & \max_{x_2 \leq x \leq 1/8} \theta(x) \\ &= \max_{x_2 \leq x \leq 1/8} \left(\frac{-x \log(x)}{\log(\gamma x^b)} \cdot \frac{\log(\gamma x^b)}{\log(1 + \gamma x^b)} \right) \\ &\leq \left(\max_{x_2 \leq x \leq 1/8} \frac{-x \log(x)}{\log(\gamma x^b)} \right) \left(\max_{x_2 \leq x \leq 1/8} \frac{\log(\gamma x^b)}{\log(1 + \gamma x^b)} \right) \\ &= \left(\theta(1/8) \frac{\log(1 + \gamma(1/8)^b)}{\log(\gamma(1/8)^b)} \right) \left(\frac{\log(\gamma(1/8)^b)}{\log(1 + \gamma(1/8)^b)} \right) \\ &= \theta(1/8) \end{aligned}$$

where we have used the fact (which can be verified using differentiation) that the function $-x \log(x)/\log(\gamma x^b)$ is strictly increasing over the interval $[x_2, 1/8]$. ■

ACKNOWLEDGMENT

We would like to thank Martin Wainwright for helpful discussions and pointers. This work was supported in part by ARO MURI No. W911NF-06-1-0076.

REFERENCES

- [1] M. Unser, "Sampling—50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 567–587, Apr. 2000.
- [2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [3] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [4] E. J. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
- [5] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. New York, NY: Springer Verlag, 1993.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. of Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
- [7] A. J. Miller, *Subset selection in regression*. New York, NY: Chapman-Hall, 1990.
- [8] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Stat.*, vol. 34, pp. 1436–1462, 2006.
- [9] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. of Machine Learning Research*, vol. 51, no. 10, pp. 2541–2563, Nov. 2006.
- [10] M. J. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting," Department of Statistics, UC Berkeley., Tech. Rep. 725, Jan. 2007.
- [11] S. Aeron, M. Zhao, and V. Saligrama, "On sensing capacity of sensor networks for the class of linear observation, fixed snr models," Jun 2007, arXiv:0704.3434v3 [cs.IT].
- [12] —, "Fundamental limits on sensing capacity for sensor networks and compressed sensing," Apr. 2008, arXiv:0804.3439v1 [cs.IT].
- [13] M. J. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
- [14] —, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inform. Theory*, vol. 55, pp. 5728–5741, Dec. 2009.
- [15] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010, (see also arXiv:0711.0366v1 [cs.IT], Nov. 2007).
- [16] G. Reeves, "Sparse signal sampling using noisy linear projections," Department of EECS, UC Berkeley, Tech. Rep. UCB/EECS-2008-3, Jan. 2008.
- [17] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.
- [18] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.
- [19] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, pp. 2967–2979, Jun. 2010.
- [20] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," Feb. 2010, arXiv:1002.4458v1 [cs.IT].
- [21] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, vol. 3, Atlanta, GA, May. 1996, pp. 1689–1692.
- [22] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
- [23] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, Jan. 2009.
- [24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [25] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
- [26] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. on Pure and Applied Math.*, vol. 59, pp. 1207–1223, Feb. 2006.
- [27] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
- [28] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
- [29] J. J. Fuchs, "Recovery of exact sparse representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 51, no. 10, pp. 3601–3608, Oct. 2005.
- [30] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
- [31] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
- [32] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1 -norm solution is also the sparsest solution," *Comm. on Pure and Applied Math.*, vol. 59, no. 6, pp. 797–829, Jun. 2006.
- [33] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
- [34] P. Massart, *Concentration Inequalities and Model Selection*. Springer, 2007.
- [35] C. Weidmann, "Oligoquantization in low-rate lossy source coding," Ph.D. dissertation, EPFL, Lausanne, Switzerland, Jul. 2000.
- [36] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *J. on Applied Signal Processing*, vol. 2006, pp. 1–19, Mar. 2006.
- [37] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep. 2006.
- [38] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Rate-distortion bounds for sparse approximation," in *Proc. IEEE Statist. Sig. Process. Workshop*, Madison, WI, Aug. 2007, pp. 254–258.
- [39] —, "Compressive sampling and lossy compression," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 48–56, Mar. 2008.
- [40] C. Weidmann and M. Vetterli, "Rate distortion behavior of sparse sources," Dec. 2008, submitted to IEEE Trans. Inform. Theory.
- [41] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.
- [42] G. Reeves and M. Gastpar, "“compressed” compressed sensing," in *Proc. IEEE Int. Symp. on Inform. Theory*, Austin, TX, Jun. 2010.
- [43] —, "A note on optimal support recovery in compressed sensing," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, Monterey, CA, Nov. 2009.
- [44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [45] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
- [46] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28(5), pp. 1302–1338, 2000.